P

P_4

Partitioned preassigned pivot procedure. A procedure for arranging the basis matrix of a linear-programming problem into as near a lower triangular form as possible. Such an arrangement helps in maintaining a sparse inverse, given that the original data set for the associated linear-programming problem is sparse.

See

- ▶ Linear Programming
- ▶ Revised Simplex Method

Packing Problem

The integer-programming problem defined as follows:

$$\begin{aligned} \text{Maximize} \quad & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{E}\boldsymbol{x} \leq \boldsymbol{e} \end{aligned}$$

where the components of $\boldsymbol{E}$ are either 1 or 0, the components of the column vector $\boldsymbol{e}$ are all ones, and the variables are restricted to be either 0 or 1. The idea of the problem is to choose among items or combinations of items that can be packed into a container and to do so in the most effective way.

See

- ▶ Bin-Packing
- ▶ Set-covering Problem
- ▶ Set-partitioning Problem

Palm Measure

▶ Markovian Arrival Process (MAP)

Parallel Computing

Jonathan Eckstein
Rutgers, The State University of New Jersey,
Livingston Campus, New Burnswick, NJ, USA

Introduction

Parallel computing is the use of a computer system that contains multiple, replicated arithmetic-logical units (ALUs), programmable to cooperate concurrently on a single task. Between 2000 and 2010, parallel computing underwent a sea change. Prior to this decade, the speed of single-processor computers advanced steadily, and parallel computing was generally employed only for applications requiring more computing power than a standard PC processor chip could deliver. Taking advantage of Moore's Law (Moore 1965), which predicts the steady increase in the number of transistors that can be packed into a given chip area, microprocessor manufacturers built processors that could execute a single stream of calculations at steadily increasing speeds. In the 2000–2010 decade, Moore's law continued to hold, but the way that chip builders used the ever-increasing number of transistors began to change. Applying ever-larger number of transistors to a single sequential stream of instructions began to encounter diminishing returns, and while smaller transistors enabled increasing clock

S.I. Gass, M.C. Fu (eds.), *Encyclopedia of Operations Research and Management Science*,
DOI 10.1007/978-1-4419-1153-7, © Springer Science+Business Media New York 2013

speeds, clock speeds are limited by energy consumption and heat dissipation issues. To use the ever-increasing number of available transistors, processor designers began placing multiple processor cores, essentially multiple processors, on each CPU chip. In the laptop and desktop markets, processors with four cores are now common, and CPU chips with only a single processing core are now rare. Thus, parallel processing is no longer only an effort to advance over the power available from mainstream computing platforms such as desktop and laptop computers; it has now become an integral part of such mainstream platforms.

Kinds of Parallel Computers

The taxonomy of Flynn (1972) classifies parallel computers as either SIMD (Single Instruction, Multiple Data) or MIMD (Multiple Instruction, Multiple Data). In SIMD architectures, a single instruction stream controls all the ALUs in a synchronous manner. In MIMD architectures, each ALU has its own instruction stream and its own instruction decoding hardware. The two approaches are not mutually exclusive: an approach sometimes called MSIMD (Multiple SIMD) combines multiple blocks of SIMD processors, with each block having its own instruction stream. There was active competition between SIMD and MIMD through the 1980s, but MIMD emerged as the clear winner in the 1990s. SIMD, however, has been staging a quiet resurgence in the form of GPUs (Graphics Processing Units), which typically have an MSIMD organization, as discussed below. Some confusion surrounds the term SIMD, as processor manufacturers also apply it to certain graphics-oriented special machine instructions that process blocks of data. These instructions are not necessarily completely parallel in the classic sense, but instead may simply take advantage of pipelining techniques to achieve higher utilization of ALU hardware than for standard scalar-operand instructions.

Another important distinction is between local and shared memory. In pure local-memory architectures, each processor has its own memory bank, and information may be moved between different processors only by messages passed through a communication network. On the other end of the spectrum are pure shared-memory designs, also called SMPs (Symmetric MultiProcessors), in which there is a single global memory bank that is equally accessible to all processors. Such designs provide performance and ease of programming for small numbers of processors, and are currently the most common, since they are used in desktop- and laptop-level multicore processor chips. In a more powerful server or workstation, two or more processor chips, each with four to six processor cores, share a single global memory. As with MIMD and SIMD, it is also possible to blend global and local memory approaches. For example, a system might be composed of dozens or hundreds of processing nodes, each node consisting of two to twelve processor cores sharing a single memory bank.

In large-scale systems without global memory, it is not generally practical to provide a dedicated connection between every pair of processors. Popular interconnection patterns include rings, grids, meshes, toroids, butterflies, and hypercubes. In academic circles, there has been an extensive debate on the merits of various interconnection topologies. However, the details of the interconnection pattern may not be critical for the kinds of parallel computers that currently exist, which generally range in size from a few processing nodes to thousands of nodes. At such scales, the critical considerations are the speed of the interconnection links, the overhead and latency associated with communication, and elementary non-interference properties. Non-interference means that sending a message from processor *A* to processor *B* should generally not interfere with processor *C* sending to processor *D*.

One way to construct a parallel computing system is simply to combine standard desktop or workstation computers, an approach known as a cluster or CoW (Cluster of Workstations). However, the local-area networks that usually connect such systems may significantly limit performance for some applications. Faster, special-purpose communication networks such as Myrinet or Infiniband may be used to improve the performance of dedicated cluster systems. Cooling and energy consumption can become significant limiting factors in constructing large CoW systems, and are also important design considerations in building higher-performance parallel supercomputers.

Another approach is to assemble ad hoc parallel systems from the background or off-hour capacity of collections of desktop computers, an approach known as grid computing, a term meant to invoke an infrastructure of computing resources resembling the electric power grid. This approach requires no special hardware, but does need specialized software such as the Condor scheduling system (Litzkow et al. 1988). Communication between instruction streams can be particularly slow in such environments, however, and algorithms must be fault tolerant, i.e., resilient to processors unpredictably disappearing from the available pool, possibly in mid-computation.

A nascent trend is GPU computing (Owens et al. 2008). The demands of ever-more sophisticated animation, driven mainly from the personal computer gaming industry, have led graphics adaptors to evolve into special-purpose parallel computing engines, often far more powerful in terms of floating point operations per second (flops) than their host processors. Modern graphic processors typically have an MSIMD structure, consisting of independent blocks of SIMD processing units. Consumer level GPUs typically contain hundreds of ALUs, at a cost of a few dollars each. In GPU computing, one uses GPU hardware for other purposes than graphics processing. Graphics processors typically have a global memory with a high bandwidth connection to their processors, but this memory is often distinct from main CPU memory.

Programming Models

The primary distinction among styles of parallel computer programming is between data-parallel and control-parallel specification of concurrency. In the data-parallel model, also called SPMD (Single Program Multiple Data), the program essentially specifies a single thread of control, but individual statements may manipulate large arrays of data in an implicitly parallel way. For example, if A, B, and C are arrays of the same size of shape, the statement $A = B + C$ might replace each element of A by the sum of the corresponding elements of B and C. Responsibility for portions of each array is typically partitioned between multiple processors, so they divide the work and perform it concurrently. Communication in data-parallel programs is typically invoked through certain standard intrinsic functions. For instance, the expression $SUM(A)$ might represent the sum, across all processors, of all A's elements, computed by whatever algorithm is optimal for the current hardware.

Data-parallel languages were originally developed for SIMD architectures, but data-parallel and SIMD are not synonymous. MIMD systems may be programmed in a data-parallel manner when it suits the application at hand. Currently, the most prevalent data-parallel programming language is High Performance FORTRAN, or HPF (Koelbel et al. 1993). HPF has its roots in FORTRAN 90 (Metcalf and Reid 1990).

In control-parallel programming, the programmer specifies a distinct thread of control for each processing unit capable of one. Often, each processing unit has the same program, but takes a completely different path through it. If shared memory is available, threads may communicate via memory, using mechanisms called locks or critical sections to prevent simultaneous or inconsistent writes to the same location. Otherwise, threads must communicate by sending and receiving messages, a style called message passing. Note that shared-memory systems may also be programmed in a message-passing style, allowing for relatively straightforward migration to larger, non-shared-memory systems. Control parallel programs are typically written in standard sequential programming languages such as C, C++, or FORTRAN, handling messages and memory interlocks via special subroutine libraries. For message passing, the principle standardized, portable subroutine libraries are based on the MPI standard (Snir et al. 1996). At least three open-source implementations of MPI are available, and system manufacturers and integrators often provide their own optimized implementations.

For shared-memory programming, common standards include Posix threads (Butenhof 1997), in which a process spawns new threads by calling special operating system routines, and OpenMP (Dagum and Menon 1998), in which parallelism is specified by special compiler directives intermixed with standard code from the underlying C, C++, or FORTRAN language. Another alternative is Cilk (Blumofe et al. 1995; Leiserson 2009), which extends the standard C and C++ languages with new parallelism-specifying syntax.

It is generally accepted that control-parallel programs are harder to analyze, understand, develop, and debug than data-parallel programs, due to complicated race and deadlock conditions that can easily develop between threads. On the other hand, the data-parallel programmer must sacrifice significant flexibility. Data parallelism is most readily applied to problems that require large, extremely regular array data structures. Irregular, sparse data structures are more the norm in operations research, and hence most of the field's successful applications of parallel computing have employed control parallelism.

Control-parallel programs can also exhibit nondeterminism: run twice on the same data, they may obtain different solutions or exhibit very different run times. Such effects occur because small differences in the timing of events may cause control-parallel programs to take complete different execution paths (serial programs that base branching decisions on measurements of clocks or timers may exhibit similar behavior). Such nondeterminism can typically be controlled and essentially eliminated, but sometimes at significant cost in performance.

Speedup, Efficiency and Scalability

If T_p is the time to solve a give problem using p processors, and T_1 is the time to solve the same problem with a single processor (using the best sequential algorithm, if it can be defined), then a key concept is speedup, defined to be $S_p = T_1/T_p$. Efficiency is then defined to be S_p/p, or, roughly speaking, the effectively used fraction of the raw computing power available. The main goal of parallel algorithm designers is to obtain *linear* speedups that grow roughly linearly with p, or, equivalently, efficiencies that do not approach 0 as p increases. In principle, speedups cannot be above linear and efficiencies cannot exceed 1; in practice, such effects can sometimes occur for specific problem instance because the "best" sequential algorithm for a particular problem is not always easily defined. In a search problem, for example, a run of a parallel algorithm might explore early in its history a portion of the search space that a standard serial implementation might not encounter until the later portions of its execution. If this portion of the search space contains the problem solution, an apparently superlinear speedup may result.

A key motivation for using parallel computing is to solve ever-larger problems. Thus, rather than concerning oneself with obtaining very large speedups for a fixed-size problem, it may be more important to study the effect on total solution time as the problem data and number of processors grow in some proportional or related way. This concept is called scalability (Kumar and Gupta 1994).

Applications in Operations Research

Parallel computing is taking an increasing role in operations research, but it has not had nearly the effect on the practice of the field as it has, for example, in computational fluid dynamics. This phenomenon is due largely to the lack of efficient parallel methods for factoring and related operations on irregularly structured sparse matrices. Such operations are essential to the sparse active set and Newton methods that form the core of operations research's numerical optimization algorithms. However, successes have been reported for specially structured problems amenable to decomposition methods, including stochastic programming — see for example Gondzio and Grothey (2007) — and on dense problems. Parallelism has also proved very useful in branch-and-bound and related search algorithms, and in a variety of randomized algorithms.

Currently, the leading vendors of linear/integer-programming software all offer some form of parallel branch-and-cut implementation for solving mixed integer programs; such implementations are typically for shared-memory systems; some are deterministic, others nondeterministic, and some offer the option of either a deterministic or nondeterministic mode. Some software vendors also offer parallel interior point linear-programming software, although speedups in pure linear programming are less dependable than for branch and bound.

Parallel open-source software for operations research operations research is becoming increasingly available. Several projects in the COIN-OR collection (Lougee-Heimer 2003) are aimed at parallel computing (typically through MPI), and several others offer the option of parallel execution.

Simulation applications with many independent trials or scenarios are also natural applications for parallel computing. A general principle seems to be that one should take advantage of problem structure to localize troublesome operations, most typically sparse matrix factorization, onto individual processors.

Another approach is to try radically new algorithms that avoid such operations completely, and are highly parallelizable. One should remember, however, that parallelism is not a panacea that can easily make inappropriate or "brute force" methods competitive.

Early references on the relationships between parallel computing and OR/MS include Barr and Hickman (1993) and Eckstein (1993).

See

- ▶ Integer and Combinatorial Optimization
- ▶ Simulation of Stochastic Discrete-Event Systems
- ▶ Stochastic Programming

References

Barr, R. S., & Hickman, B. L. (1993). Reporting computational experiments with parallel algorithms: Issues, measures and experts = opinions. *ORSA Journal of Computing, 5*, 2–18.

Bertsekas, D. P., & Tsitsiklis, J. (1989). *Parallel and distributed computation: Numerical methods*. Englewood Cliffs, NJ: Prentice-Hall.

Blumofe, R. D., Joerg, C. F., Kuszmaul, B. C., Leiserson, C. E., Randall, K. H., Zhou, Y. (1995). Cilk: An efficient multithreaded runtime system. *Proceedings of the Fifth ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, Santa Barbara, California, 207–216.

Butenhof, D. R. (1997). *Programming with Posix threads*. Boston, MA: Addison-Wesley.

Dagum, L., & Menon, R. (1998). OpenMP: An industry standard API for shared-memory programming. *IEEE Computational Science and Engineering, 5*, 46–55.

Eckstein, J. (1993). Large-scale parallel computing, optimization, and operations research: A survey. *ORSA Computer Science Technical Section Newsletter*, 14(2), 1, 8–12.

Flynn, M. J. (1972). Some computer organizations and their effectiveness. *IEEE Transactions on Computers, C-21*, 948–960.

Gondzio, J., & Grothey, A. (2007). Parallel interior-point solver for structured quadratic programs: Application to financial planning problems. *Annals of Operations Research, 152*, 319–339.

Kindervater, G. A. P., & Lenstra, J. K. (1988). Parallel computing in combinatorial optimization. *Annals of Operations Research, 14*, 245–289.

Koelbel, C. H., Loveman, D. B., Schreiber, R. S., Steele, G. L., Zosel, M. E. (1993). *The high performance Fortran handbook*. Cambridge, MA: MIT Press.

Kumar, V., & Gupta, A. (1994). Analyzing scalability of parallel algorithms and architectures. *Journal of Parallel and Distributed Computing, 22*, 379–391.

Leighton, F. T. (1991). *Introduction to parallel algorithms and architectures: Arrays, trees, and hypercubes*. San Mateo, CA: Morgan Kaufmann.

Leiserson, C. E. (2009). The CILK++ concurrency platform. *Proceedings of the 46th Annual Design Automation Conference*, ACM, San Francisco, California, 522–527.

Litzkow, M. J., Livny, M., & Mutka, M. W. (1988). Condor-a hunter of idle workstations. *Proceedings of the 8th International Conference on Distributed Computing Systems*, IEEE, San Jose, California, 104–111.

Lougee-Heimer, R. (2003). The common optimization interface for operations research: Promoting open-source software in the operations research community. *IBM Journal of Research and Development, 47*, 57–66.

Metcalf, M., & Reid, J. (1990). *Fortran 90 explained*. Oxford, UK: Oxford University Press.

Moore, G. (1965). Cramming more components onto integrated circuits. *Electronics, 38*(8), 114–117.

Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E., Phillips, J. C. (2008). GPU computing. *Proceedings IEEE*, 96, 879–899.

Snir, M., Otto, S. W., Huss-Lederman, S., Dongarra, J., Kowalik, J. S. (1996). *MPI: The complete reference*. Cambridge, MA: MIT Press.

Zenios, S. A. (1994). Parallel and supercomputing in the practice of management science. *Interfaces, 24*, 122–140.

Parameter

A quantity appearing in a mathematical model that is subject to controls beyond those affecting the decision variables.

Parameter-Homogeneous Stochastic Process

A stochastic process in which distribution properties between the two index parameter points t_1 and t_2, $t_1 \leq t_2$, depend only on the difference $t_2 - t_1$, and not on the specific values of t_1 and t_2. In the many applications where the parameter set is time, whether discrete or continuous, it is called a time-homogeneous stochastic process.

Parametric Bound

An optimal value function or solution point bound as a function of problem parameters.

Parametric Linear Programming

In the general linear-programming problem of

$$\begin{aligned} \text{Minimize} \quad & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq 0 \end{aligned}$$

it is often appropriate to study how the optimal solution changes when some of the data are functions of a single parameter λ. Most mathematical programming systems allow parametric analysis of the cost coefficients (PAROBJ), the right-hand-side elements (PARARHS), joint analysis of the objective function and right-hand-side elements (PARARIM), and the parametric analysis of the data in a row (PARAROW).

Parametric Programming

Tomas Gal
Fern Universität in Hagen, Hagen, Germany

Introduction

The meaning of a parameter as used here is best explained by a simple example. Recall that a parabola can be expressed as follows: $y = ax^2$, $a \neq 0$. Setting $a = 1$, a parabola is obtained that has a different shape from the parabola when setting, for example, $a = 5$. In both cases, however, there are parabolas that obey specific relationships; only the shapes are different. Hence, the parabola $y = ax^2$ describes a family of parabolas and the parameter a specifies the shape.

Consider the general mathematical-programming problem:

$$\text{Max } z = f(\boldsymbol{x}) \tag{1}$$

$$\text{subject to } g(\boldsymbol{x}) \leq 0 \tag{2}$$

Introducing one or more parameters into f or g, the model stays the same, but for each value of the parameter(s) one obtains a specific problem.

In setting up a mathematical optimization model, one of the first tasks is to collect data. The collected data might, however, be inaccurate, be of a stochastic character, be uncertain or be deficient in other ways. Therefore, it is appropriate to introduce parameters that enable to analyze the influence of specific data elements on the optimal solution. This can be done by:

1. Introducing the parameter(s) at the beginning when setting up the model, or
2. Introducing the parameter(s) after an optimal solution has been found.

The latter case is called postoptimal analysis (POA) and is applied much more frequently than the first case.

Postoptimal analysis is a very important tool that should be used in the framework of a good report generator (Gal 1993). The corresponding decision maker (DM) would then have information with which the DM can select a firm optimum. POA consists of several analyses, the most important of which is sensitivity analysis (SA). A sort of extended SA is parametric programming (PP). In nonlinear programming, SA corresponds to perturbation analysis, in which, after having found an optimal solution, some of the initial data are perturbed and the influence of the perturbation on the outcome is analyzed (Drud and Lasdon 1997).

Historical Sketch

Advanced methods for SA and PP for linear programming have been developed. In the 1950s, Orchard-Hays (in his master's thesis), Manne (1953), Saaty and Gass (1954), Gass and Saaty (1955) published the first works on parametric programming. By the end of the 1960s, the first monograph on parametric programming appeared (Dinkelbach 1969), followed by the monograph and book by Gal (1973, 1979). In 1979, the first Symposium on Data Perturbation and Parametric Programming was organized by A.V. Fiacco in Washington, D.C., with such a symposium being held every year since. (From 1999, Adi Ben Israel has been the organizer). Several monographs (Bank et al. 1982; Guddat et al. 1991) and special journal issues have been published in the 1970s and 1980s. More details on the history of PP are given in Gal (1980, 1983). A bibliography with over 1,000 items is given in Gal (1994b); see also Gal and Greenberg (1997).

Postoptimal Analysis

Assume that the mathematical optimization model under consideration is a linear program of the form:

$$\text{Max } z = \boldsymbol{c}^T\mathbf{x} \tag{3}$$

$$\text{subject to } \boldsymbol{Ax} = \boldsymbol{b}, \quad \boldsymbol{x} \geq \mathbf{0} \tag{4}$$

where $\boldsymbol{c}$ is an n-vector of objective function coefficients (OFC) c_j, x is an n-vector of the decision variables x_j, $\boldsymbol{A}$ is an $m \times n$ matrix of the technological coefficients a_{ij}, $m < n$, $\boldsymbol{b}$ is an m vector of the right-hand-side (RHS) elements b_i. All vectors are column vectors.

Suppose that the problem defined by (3) and (4) has an optimal basic feasible solution $x_{\boldsymbol{B}} = \boldsymbol{B}^{-1}\,\boldsymbol{b}$, where $\boldsymbol{B}^{-1}$ is the inverse of the $m \times m$ basic matrix $\boldsymbol{B}$ (the basis) consisting of m linearly independent columns of $\boldsymbol{A}$. Here, $x_{\boldsymbol{B}}$ is an m-dimensional solution vector. This means that the following solution elements and simplex method elements are determined:

1. The maximal value of the objective function (OF), $z_{\max}$,
2. The values of the basic variables x_i, $i = 1,\ldots, m$, and
3. The reduced costs $d_j = z_j - c_j$, $j = 1,\ldots, n$.

In the framework of POA, an evaluation of the above solution elements is to be performed. This means that the DM is provided with information about the meaning of the values of the basic variables, the DM is told which resources are used and are critical (values of slack variables), and interpret the values of the opportunity costs and shadow prices. It is also possible to carry out a suboptimal analysis, that is, show the DM what happens if one or several nonbasic variables were introduced into the solution at a positive level.

Sensitivity Analysis

The POA would continue by performing a SA with respect to the OF and the RHS. This analysis is usually a part of the solution output for just about all linear-programming software. It is called OFC-ranging and RHS-ranging, respectively. Behind such analyses is the introduction of a scalar parameter, t or λ, in the form

$$c_j(t) = c_j + t,\ j \text{ fixed} \tag{5}$$

or

$$b_i(\lambda) = b_i + \lambda,\ \ i \text{ fixed} \tag{6}$$

SA finds a critical interval T_j or Λ_i, such that for all $t \in T_j$ or $\lambda \in \Lambda_i$, respectively, the (found) optimal basis $\boldsymbol{B}$ remains the same (so called optimal basis invariancy. For other kinds of invariancies see, e.g., Hladik 2010; Hadigheh et al. 2007). The critical values, that is, the upper and lower bounds of the critical interval can be easily determined by certain formulas (Gass 1985). A change in a RHS element b_i causes, in general, the values of the basic variables and the value of $z_{\max}$ to change, while a change in an OFC c_j causes, in general, the values of the reduced costs and the value of $z_{\max}$ to change. Such information is of great value to the DM. An assumption of this type of SA is that to investigate how the optimal solution would vary with respect to a change in one data element, while holding all other data fixed. Analysis of multiple changes can be done in a limited manner by the techniques of the hundred percent rule (Bradley et al. 1977) and tolerance analysis (Ashram 2007; Filippi 2005; Hladik 2008a, b; Wendell 1985, 2004).

Parametric Analysis

For an element b_i of the RHS, the question is asked: for what range of values of the parameter λ in (6) does there exist an optimal solution to (3) and (4)? Given such values, one can move from the original optimal basis and generate a sequence of optimal bases, with each basis associated with a critical interval of the parameter. Such an analysis provides the DM with a full range of possible solutions from which a subset of optimal solutions appropriate for the given problem can be selected. The DM then chooses a certain value of the parameter and, thus, a corresponding optimal solution for the parametric range of b_i (λ).

Note that a similar analysis can be performed with respect to the parametric OFC, as given by (5). Moreover, taking into account the possibility that a parameter introduced in the RHS may influence some (or several) OFC or vice versa, it is possible to perform a RIM parametric analysis, that is, find a sequence of optimal bases to each of which a critical interval for the RHS-and for the OFC-parameters are associated simultaneously. Standard RHS, OFC and RIM parametric analysis procedures are usually included in linear-programming software.

It is also possible to perform a sensitivity or parametric analysis with respect to the elements a_{ij} of the matrix $\boldsymbol{A}$. The corresponding procedures are, unfortunately, not incorporated into linear-programming software as the underlying formulas are a bit too complex. However, some software enables one to compute a series of linear programs in each of which slightly changed values of the $\{a_{ij}\}$ are chosen.

Up to now, the simplest parametric case having one parameter with a coefficient equal to 1 has been discussed. The above cases can, however, also be carried out when:

(i) A scalar parameter is introduced into several elements of the RHS and/or OFC with coefficients which differ from 1, and

(ii) A parameter-vector (vector of parameters) is introduced into several elements of the RHS and/or OFC with their respective coefficients different from 1.

As far as case (i) is concerned, to each optimal basis a critical interval is associated. In case (ii), each optimal basis is associated with a higher dimensional convex polyhedral set of parameters. In the RIM case, each optimal basis is associated with a higher dimensional interval, a box, provided that the parameters in the RHS and OFC are independent from each other. The larger the number of parameters in the parameter-vector, the more difficult it is to interpret the results and for the DM to find an appropriate optimal basis. In such cases, an interactive approach is recommended in which the parametric specialist helps the DM to select an appropriate solution.

Applications

There are two kinds of uses of PP:

1. Introducing parameters into various classes of mathematical-programming problems for solving these problems via parameterization; and
2. Practical applications.

As to (1), the introduction of parameters helps to solve problems from the areas of nonconcave mathematical programming, decomposition, approximation, and integer programming. Also, note that by replacing the OFC in (3) and (4) with a matrix $\boldsymbol{C}$ times a parameter-vector $\boldsymbol{t}$ the following problem is obtained

$$\begin{aligned}&\text{Max } z = (\boldsymbol{C}^T\boldsymbol{t})\boldsymbol{x},\\ &\text{subject to} \quad \boldsymbol{Ax} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\end{aligned}$$

which is a scalarized version of a linear multiobjective-programming problem (Steuer 1986). Methods for solving the corresponding homogeneous multi-parameter-programming problem provide a procedure to determine the set of all efficient solutions of the corresponding multiobjective problem (Gal 1994b).

As to (2), SA and/or PP has been used in the pipeline industry, in capital budgeting, for farm decision making, refinery operations, for return maximization in an enterprise, and a number of other applications (Gal 1994b).

SA and PP in Other Fields

Theoretical and methodological works have been published about SA and/or PP in linear and nonlinear complementarity problems, control of dynamic systems, fractional programming, geometric programming, integer and quadratic programming problems, transportation problems. A more detailed survey with corresponding references is given in Gal (1994b) (1988), see also, e.g., Ravi and Wendell (1988), Hladik (2008b), Dawande and Hooker (2000), Faisca et al. (2009), Kheirfam (2010).

Degeneracy

Recall that a basic feasible solution to a linear-programming problem is called primal degenerate when at least one element of this solution equals zero. The corresponding extreme point of the feasible set, that is, of the convex polyhedron, is then also called degenerate. Degeneracy causes various kinds of efficiency and convergence problems and special precautions must be taken when performing SA for a degenerate extreme point. Degeneracy influences even POA, especially the determination of opportunity costs and shadow prices. When performing SA, the main rule – determining the critical interval such that the original optimal basis does not change – is no longer valid because for a degenerate solution many bases are associated with it. A theoretical discussion of this problem is given in Kruse (1986), a bibliography is found in Gal (1994a). Note that standard software analysis for RHS-or OFC-ranging yield false results when degeneracy is involved.

Concluding Remarks

For linear programming and related mathematical areas, SA and PP have become important tools for analyzing variations in initial data, for obtaining better insight into and gaining more information about the related mathematical model, for improving understanding of model building in general, and as aids in solving a wide range of mathematical problems.

See

- ▶ Degeneracy
- ▶ Degeneracy Graphs
- ▶ Linear Programming
- ▶ Multiobjective Programming
- ▶ Perturbation Methods
- ▶ Sensitivity Analysis

References

Ashram, H. (2007). Construction of the largest sensitivity region for general linear programs. *Applied Mathematics and Computation, 189*, 1435–1447.

Bank, B., Guddat, J., Klatte, D., Kummer, B., & Tammer, T. (1982). *Nonlinear parametric optimization*. Berlin: Akademie Verlag.

Bradley, S. P., Hax, A. C., & Magnanti, T. L. (1977). *Applied mathematical programming*. Reading, MA: Addison-Wesley.

Dawande, M. W., & Hooker, J. N. (2000). Inference-based sensitivity analysis for mixed integer/linear programming. *Operations Research, 48*, 623–634.

Dinkelbach, W. (1969). *Sensitivitätsanalysen und parametrische Programmierung*. Berlin: Springer Verlag.

Drud, A. S., & Lasdon, L. (1997). Nonlinear programming. In T. Gal & H. J. Greenberg (Eds.), *Advances in sensitivity analysis and parametric programming*. Norwell, MA: Kluwer.

Faisca, N. P., Kosmidis, V. D., Rustem, B., & Pistikopoulos, E. N. (2009). Global optimization of multi-parametric MILP problems. *Journal Global Optimization, 45*(1), 131–151.

Filippi, C. (2005). A fresh view on the tolerance approach to sensitivity analysis in linear programming. *European Journal of Operational Research, 167*, 1–19.

Gal, T. (1973). *Betriebliche Entscheidungsprobleme, Sensitivitätsanalyse und parametrische Programmierung*. Berlin: W. de Gruyter.

Gal, T. (1979). *Postoptimal analyses, parametric programming and related topics*. New York: McGraw Hill.

Gal, T. (1980). A 'historiogramme' of parametric programming. *Journal of the Operational Research Society, 31*, 449–451.

Gal, T. (1983). A note on the history of parametric programming. *Journal of the Operational Research Society, 34*, 162–163.

Gal, T. (1993). Putting the LP survey into perspective. *OR/MS Today, 19*(6), 93.

Gal, T. (1994a). Selected bibliography on degeneracy. *Annals Operations Research.*

Gal, T. (1994b). *Postoptimal analyses and parametric programming*. Berlin: W. de Gruyter. Revised and updated edition.

Gal, T., & Greenberg, H. J. (Eds.). (1997). *Advances in sensitivity analysis and parametric programming*. Norwell, MA: Kluwer.

Greenberg, H. J. (1993). *A computer-assisted analysis system for mathematical programming models and solutions: A user's guide for ANALYZE*. Norwell, MA: Kluwer.

Gass, S. I. (1985). *Linear programming* (5th ed.). New York: McGraw-Hill.

Gass, S. I., & Saaty, T. L. (1955). The parametric objective function. *Naval Research Logistics Quarterly, 2*, 39–45.

Guddat, J., Guerra Vazquez, F., & Jongen, H. T. (1991). *Parametric optimization: Singularities, path following and jumps*. Stuttgart/New York: B. G. Teubner/Wiley.

Hadigheh, A. G., Mirnia, K., & Terlaky, T. (2007). Active constraint set invariancy sensitivity analysis in linear optimization. *JOTA, 133*, 303–315.

Hladik, M. (2008a). Additive and multiplicative tolerance in multiobjective linear programming. *Operations Research Letters, 36*, 393–396.

Hladik, M. (2008b). Computing the tolerance in multiobjective linear programming. *Optimization Methods and Software, 23*, 731–739.

Hladik, M. (2010). Multiparametric linear programming: Support set and optimal partition invariancy. *European Journal of Operational Research, 202*, 25–31.

Kheirfam, B. (2010). Sensitivity analysis in multi-parametric strictly convex quadratic optimization. *Matem. Vesnik, 62*, 95–107.

Kruse, H.-J. (1986). *Degeneracy graphs and the neighborhood problem* (Lecture Notes in economics and mathematical systems No. 260). Berlin: Springer Verlag.

Manne, A. S. (1953). *Notes on parametric linear programming, RAND Report P-468*. Santa Monica, CA: The Rand Corporation.

Ravi, N., & Wendell, R. E. (1988). Tolerance approach to sensitivity analysis in network linear programming. *Networks, 18*, 159–181.

Saaty, T. L., & Gass, S. I. (1954). The parametric objective function, Part I. *Operations Research, 2*, 316–319.
Steuer, R. E. (1986). *Multiple criteria optimization: Theory, computation, and application*. New York: Wiley.
Wendell, R. E. (1985). The tolerance approach to sensitivity analysis in linear programming. *Management Science, 31*, 564–578.
Wendell, R. E. (2004). Tolerance sensitivity and optimality bounds in linear programming. *Management Science, 50*, 797–803.

Parametric Solution

A solution expressed as a function of problem parameters.

Pareto-Optimal Solution

If a feasible deviation from a solution to a multiobjective problem causes one of the objectives to improve while some other objective degrades, the solution is termed a Pareto-optimal. Such a solution is also called an efficient or nondominated solution.

See

- ▶ Efficient Solution

Partial Balance Equations

In Markov chain models of queueing networks, a subset of the global balance equations that may be satisfied at a node (station), i.e., a balance of mean flow rates or probability flux. Also known as local balance equations, falling between global balance equations and detailed balance equations.

See

- ▶ Detailed Balance Equations
- ▶ Global Balance Equations
- ▶ Markov Chains
- ▶ Networks of Queues
- ▶ Queueing Theory

Partial Pricing

When determining a new variable to enter the basis by the simplex method, it is somewhat computationally inefficient to price out all nonbasic columns, as is the way of the standard simplex algorithm or its multiple pricing refinement. The scheme of partial pricing starts by searching the nonbasic variables in index order until a set of candidate vectors has been found. These vectors are then used as possible vectors to enter the basis, as is done in multiple pricing. After the candidate set is depleted, another set is found by searching the nonbasic vectors from the point where the first set stopped its search. The process continues in this manner by searching and selecting candidate sets until the optimal solution is found. Although the total number of iterations to solve a problem usually increases, computational time is saved by this type of pricing strategy.

See

- ▶ Simplex Method (Algorithm)

Partially Observed Markov Decision Processes

A Markov decision process (MDP) in which the state of the system cannot be fully or precisely observed, e.g., only part of the state is known and/or the state observation has some error. In principle, such a model can be converted to a fully observed MDP by introducing an “information” or “belief” state that may be infinite dimensional, corresponding to a probability distribution over the original state.

See

- ▶ Dynamic Programming
- ▶ Markov Decision Processes

Particle Swarm Optimization

A population-based search approach for global optimization based on ideas from animal flocking.

See

- ▶ Ant Colony Optimization
- ▶ Metaheuristics
- ▶ Swarm Intelligence

References

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of IEEE International Conference on Neural Networks*, Vol. IV, pp. 1942–1948.

PASTA

Poisson Arrivals See Time Averages.

For a Poisson arrival process, the (limiting) fraction of arrivals that find (see) a process in some state equals the (limiting) overall fraction of time that the process is in that state (Wolff 1982, 1990).

See

- ▶ Poisson Arrivals

References

Wolff, R. W. (1982). Poisson arrivals see time averages. *Operations Research, 30*, 223–231.

Wolff, R. W. (1990). A note on PASTA and anti-PASTA for continuous-time Markov chains. *Operations Research, 38*, 176–177.

Path

A path in a network is a sequence of nodes and arcs that connect a designated initial node to a designated terminal node.

See

- ▶ Chain
- ▶ Cycle

Payoff Function

In a game, the mapping from the players' strategies (decisions, actions) to the gains and losses they receive. In a two-person finite action game, the payoff function is often depicted in the form of a matrix, with a single number for each matrix element in a zero-sum game.

In financial engineering, the mapping from the underlying asset(s) to the payout of a contingent claim or financial derivative.

See

- ▶ Financial Engineering
- ▶ Game Theory

Payoff Matrix

For a zero-sum, two-person game, the payoff matrix is an $m \times n$ matrix of real numbers with the entry a_{ij} representing the payoff to the maximizing player if the maximizing player plays strategy i and the minimizing player plays strategy j.

See

- ▶ Game Theory

PDA

Parametric decomposition approach.

See

- ▶ Production Management

PDF

Probability density function.

PDSA

Plan, do, study, act.

See

- ▶ Total Quality Management

Periodic Review

A type of inventory control policy in which the inventory position is assessed at the end of each of a prescribed number of discrete time periods, in contrast with continuous review, where the inventory position is monitored continuously so that orders can be placed at any time.

See

- ▶ Inventory Modeling

PERT

Program evaluation and review technique; an event-oriented, project-network diagramming technique used for planning and scheduling.

See

- ▶ Network Planning
- ▶ Program Evaluation and Review Technique (PERT)
- ▶ Project Management
- ▶ Research and Development

Perturbation

A change in a parameter, function or set.

Perturbation Analysis

Michael C. Fu
University of Maryland, College Park, MD, USA

Introduction

Perturbation analysis (PA) is a sample path technique for analyzing changes in performance measures of stochastic systems due to changes in system parameters. In terms of stochastic simulation, which is the main setting for PA, the objective is to estimate sensitivities of the performance measures of interest with respect to system parameters, preferably without the need for additional simulation runs over what is required to estimate the system performance itself. The primary application is gradient estimation during the simulation of discrete-event systems, e.g., queueing and inventory systems. Besides their importance in sensitivity analysis, these gradient estimators are a critical component in gradient-based simulation optimization methods.

Let $l(\theta)$ be a performance measure of interest with parameter (possibly vector) of interest θ, focusing on those systems where $l(\theta)$ cannot be easily obtained through analytical means and therefore must be estimated from sample paths, e.g., via stochastic simulation. Denote by $L(\theta, \omega)$ the sample performance obtained from a sample path realization ω such that $l(\theta) = E[L(\theta, \omega)]$. Although the assumption here is that the performance measure is an expectation, PA has also been applied more recently to quantiles (Hong 2009; Fu et al. 2009). The goal of PA is to efficiently estimate the effects on l of a perturbation $\theta \to \theta + \Delta\theta$, using information from a sample path ω at θ. PA addresses two different types of problems:

- $\Delta\theta \to 0$: estimating the gradient $\nabla l(\theta)$, when l is differentiable in θ.
- $\Delta\theta \neq 0$: estimating changes due to a finite perturbation, i.e., $l(\theta + \Delta\theta)$.

In the former case, no perturbation is ever actually introduced into the system (or simulation), although the idea of a perturbation may be employed as a heuristic tool in preliminary analysis.

Brief Taxonomy

To sort out the abundance of acronyms in the PA field, a brief definition of each corresponding approach is provided here, accompanied with at least one reference. Among gradient estimation techniques, the most well-known is infinitesimal perturbation analysis (IPA), which simply uses the sample derivative $dL/d\theta$ to estimate $dl/d\theta$. It is straightforward to implement and very computationally efficient; however, as shall be discussed shortly in more detail, its applicability is not universal. The books by Ho and Cao (1991), Glasserman (1991), and Cao (1994) cover IPA in detail. A very general and well-developed extension of IPA is smoothed perturbation analysis (SPA), based on the ideas of conditional expectation (Gong and Ho 1987) Although its applicability is quite broad, its implementation is usually very problem dependent. The book by Fu and Hu (1997) covers this method in full generality. Other gradient estimation techniques include rare perturbation analysis (RPA), originally based on the thinning of point processes (Brémaud and Vázquez-Abad 1992); structural IPA (SIPA), dealing specifically with structural parameters (Dai and Ho 1995); discontinuous perturbation analysis (DPA), based on the use of generalized functions (the Dirac-delta function) to model discontinuities in the sample performance function (Shi 1996); and augmented IPA (APA), another extension of IPA different from SPA (Gaivoronski et al. 1992). Techniques to estimate the effect of a finite perturbation in the parameter include finite perturbation analysis (FPA) – Ho et al. (1983); extended perturbation analysis (EPA) – Ho and Li (1988); and the augmented chain method–Cassandras and Strickland (1989). A related technique is the standard clock (SC) method, based on the uniformization of Markov chains (Vakili 1991). The books by Ho and Cao (1991) and Cassandras and Lafortune (2008) provide further references. This entry focuses on the gradient estimation techniques IPA and SPA, the most well-known and developed of the PA techniques.

Infinitesimal Perturbation Analysis

The applicability of IPA is illustrated through the use of some simple examples, at the same time contrasting the approach with the likelihood ratio/score function (LR/SF) and weak derivative (WD) estimators. Consider first the expectation of a single positive random variable X, written in two forms:

$$\begin{aligned} E[X] &= \int_0^\infty x f(x;\theta)dx \\ &= \int_0^1 X(\theta;u)du, \end{aligned}$$

where f is the PDF of X. In the first interpretation, the parameter appears inside the density, whereas in the second interpretation it appears inside the random variable defined on an underlying $U(0,1)$ random number. For example, the latter could be the inverse transform $X = F^{-1}$, where F is the CDF of X.

Differentiating $E[X]$, assuming the interchange of expectation and differentiation is permissible (via the dominated convergence theorem),

$$\frac{dE[X]}{d\theta} = \int_0^\infty x \frac{df(x;\theta)}{d\theta} dx \tag{1}$$

$$= \int_0^1 \frac{dX(\theta;u)}{d\theta} du. \tag{2}$$

Notice, however, that the conditions for the exchange will be quite different for the two interpretations. In the first interpretation, corresponding to the LR/SF and WD estimators, the conditions will be placed on the underlying density; in the case of discrete-event stochastic simulation, this means the input distributions. Since the input distributions must be known in order to perform the simulation, it is relatively easy to check the conditions. In the second interpretation, corresponding to PA estimators, the conditions will be placed on the sample performance function that is usually defined on an output stochastic process of the system.

As an example, consider an exponential random variable X with mean θ. Then $E[X] = \theta$ and $dE[X]/d\theta = 1$. The respective PDF and one random variable representation are given by

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta} e^{-x/\theta} 1\{x > 0\}, \\ X(\theta;u) &= -\theta \ln u, \end{aligned}$$

where $1\{\cdot\}$ denotes the indicator function. Differentiating,

$$\begin{aligned}\frac{df(x;\theta)}{d\theta} &= \left[\frac{x}{\theta^2}\frac{1}{\theta}e^{-x/\theta} - \frac{1}{\theta^2}e^{-x/\theta}\right]1\{x>0\}\\ &= f(x;\theta)\left[\frac{x}{\theta^2}-\frac{1}{\theta}\right]\\ &= \frac{1}{\theta e}\left[\frac{e}{\theta}\left(1-\frac{x}{\theta}\right)e^{-x/\theta}1\{0<x\le\theta\}\right.\\ &\quad\left.-\frac{e}{\theta}\left(\frac{x}{\theta}-1\right)e^{-x/\theta}1\{x>\theta\}\right],\\ \frac{dX(\theta;u)}{d\theta} &= -\ln u = \frac{X(\theta;u)}{\theta}.\end{aligned}$$

The last expression for the derivative of the density (which is itself *not* a density) expresses the quantity as the difference of two densities multiplied by a constant, known as a weak derivative representation; Fu (2006, 2008) for references. Substituting each of the three expressions into the corresponding equations (1) or (2), yields three unbiased derivative estimators:

$$\begin{aligned}\mathrm{LR/SF}: &\frac{X}{\theta}\left(\frac{X}{\theta}-1\right),\\ \mathrm{WD}: &\frac{1}{\theta e}\left[X^{(2)}-X^{(1)}\right],\\ \mathrm{IPA}: &\frac{X}{\theta},\end{aligned}$$

where $X^{(1)}$ and $X^{(2)}$ are random variables with PDFs $\frac{e}{\theta}\left(\frac{x}{\theta}-1\right)e^{-x/\theta},\ x>\theta,$ and $\frac{e}{\theta}\left(1-\frac{x}{\theta}\right)e^{-x/\theta},\ 0<x\le\theta,$ respectively.

Extending to a function of the underlying random variable,

$$\begin{aligned}\frac{dE[L(X)]}{d\theta} &= \int_0^\infty L(x)\frac{df(x;\theta)}{d\theta}dx\\ &= \int_0^1 \frac{dL}{dX}\frac{dX(\theta;u)}{d\theta}du.\end{aligned}$$

The conditions for interchanging expectation and differentiation are unaltered when differentiating the underlying density, since that portion remains unchanged, whereas they are more involved for the sample path derivative. Basically, for the chain rule to be applicable requires some sort of continuity to hold for the sample performance function with respect to the underlying random variable. This translates into requirements on the form of the performance measure and on the dynamics of the underlying stochastic system such that the interchange

$$\frac{dE[L]}{d\theta} = E\left[\frac{dL}{d\theta}\right] \tag{3}$$

holds. Roughly speaking, sample pathwise continuity of L with respect to θ will result in the interchange being valid. An important structural condition for determining the applicability of IPA for general discrete-event systems modeled as generalized semi-Markov processes is the commuting condition (Glasserman 1991).

Smoothed Perturbation Analysis

The main idea of smoothed perturbation analysis (SPA) is to use conditional expectation to smooth out discontinuities in L that cause IPA to fail. This is achieved by selecting a set of sample path quantities $\mathcal{Z}$, called the characterization, such that $E[L|\mathcal{Z}]$ – as opposed to L itself – will satisfy the interchange in (3):

$$\frac{dE[E[L|\mathcal{Z}]]}{d\theta} = E\left[\frac{dE[L|\mathcal{Z}]}{d\theta}\right].$$

Applying SPA is analogous to the variance reduction technique of conditional Monte Carlo, consisting of two main steps: choosing an appropriate $\mathcal{Z}$ and calculating $dE[L|\mathcal{Z}]/d\theta$. For generalized semi-Markov processes, as well as for other stochastic systems, this is fully explored in Fu and Hu (1997).

Queueing Example

IPA and SPA estimators are illustrated for a single-server, first come, first-served (FCFS) queue. Let A_n be the interarrival time between the $(n-1)$th and nth customer (i.i.d. with PDF f_1 and CDF F_1), X_n the service time of the nth customer (i.i.d. with PDF f_2 and CDF F_2), and T_n the system time (in queue plus in service) of the nth customer. Consider the case where θ is a parameter in the service time distribution, and the sample performance of interest is the average system time over the first N customers $\overline{T}_N = \frac{1}{N}\sum_{n=1}^{N} T_n$. The system time of a customer for a FCFS single-server queue satisfies the well-known recursive Lindley equation:

$$T_{n+1} = X_{n+1} + (T_n - A_{n+1})^+. \qquad (4)$$

The IPA estimator is obtained by differentiating (4):

$$\frac{dT_{n+1}}{d\theta} = \frac{dX_{n+1}}{d\theta} + \frac{dT_n}{d\theta} 1\{T_n \geq A_{n+1}\}, \qquad (5)$$

where

$$\frac{dX}{d\theta} = -\frac{dF_2(X;\theta)/d\theta}{dF_2(X;\theta)/dX}.$$

For example, for scale parameters, such as if θ is the mean of an exponential distribution, $dX/d\theta = X/\theta$. Using the above recursion, the IPA estimator for the derivative of average system time is given by

$$\begin{aligned}\frac{d\overline{T}_N}{d\theta} &= \frac{1}{N}\sum_{n=1}^{N}\frac{dT_n}{d\theta} \\ &= \frac{1}{N}\sum_{m=1}^{M}\sum_{i=n_{m-1}+1}^{n_m}\sum_{j=n_{m-1}+1}^{i}\frac{dX_j}{d\theta},\end{aligned} \qquad (6)$$

where M is the number of busy periods observed and n_m is the index of the last customer served in the mth busy period ($n_0 = 0$). Implementation of the estimator involves keeping track of two running quantities, one for (5) and another for the summation in (6); thus, the additional computational overhead is minimal, and *no alteration of the underlying simulation is required.* IPA is also applicable to multi-server queues and Jackson-like queueing networks (Jackson networks without the exponential distribution assumptions).

The implicit assumption used in deriving an IPA estimator is that small changes in the parameter will result in small changes in the sample performance. For example, small changes in the interarrival and service times lead to small changes in system times, as can be seen by the Lindley equation (4), but can lead to large changes in the derivative given by (5), due to the indicator function. In general, the interchange (3) will hold if the sample performance is continuous with respect to the parameter. For the Lindley equation, although T_{n+1} in (4) has a kink at $T_n = A_{n+1}$, it is still continuous at that point, which explains why IPA works. Unfortunately, the kink means that the derivative given by (5) has a discontinuity at $T_n = A_{n+1}$, so that IPA will fail for the second derivative.

For the FCFS single-server queue, SPA can be used to derive the following estimator for the second derivative of mean system time:

$$\begin{aligned}\left(\frac{d^2\overline{T}_N}{d\theta^2}\right)_{SPA} &= \frac{1}{N}\sum_{m=1}^{M}\sum_{i=n_{m-1}+1}^{n_m}\sum_{j=n_{m-1}+1}^{i}\frac{d^2X_j}{d\theta^2} \\ &+ \frac{1}{M}\sum_{m=1}^{M}\frac{f_1(T_{n_m})}{1-F_1(T_{n_m})}\left(\sum_{i=n_{m-1}+1}^{n_m}\frac{dX_i}{d\theta}\right)^2,\end{aligned}$$

where $d^2X/d\theta^2$ is well-defined when $F_2(X;\theta)$ is twice differentiable.

Inventory Example

IPA and SPA estimators are illustrated for a single-item periodic review (s, S) inventory system, in which once every period the inventory level is reviewed and, if necessary, orders are placed to replenish depleted inventory. An (s, S) ordering policy specifies that an order be placed when the level of inventory on hand plus that on order (known as inventory position) falls below the level s, and that the amount of the order be the difference between S and the present inventory position, i.e., order amounts are placed "up to S." For average inventory as the performance measure of interest, derivative estimators with respect to the policy parameters s and $q = S - s$ are provided. Note that the parameters in this example are structural, as opposed to distributional in the previous queueing example.

In the model considered, all excess demand is backlogged and eventually filled, and orders are immediately received (zero lead time), so that. inventory level and inventory position coincide. At the end of a period, demand is satisfied *before* the order placement decision is made. Let D_n be the demand in period n (i.i.d. with PDF f and CDF F), and V_n be the inventory level in period n after demand satisfaction. This quantity satisfies a recursive equation somewhat analogous to the Lindley equation:

$$V_{n+1} = \begin{cases} V_n - D_{n+1} & \text{if } V_n \geq s, \\ S - D_{n+1} & \text{if } V_n < s. \end{cases} \tag{7}$$

The sample performance is the average inventory level over N periods given by $\overline{V}_N = \frac{1}{N}\sum_{n=1}^{N} V_n$.

From a sample path point of view, the key discrete event in the system is the ordering decision each period. A change in s, with q held fixed, has no effect on these decisions, so infinitesimal perturbations in s result in infinitesimal changes in the inventory level, and hence in the sample performance function $\overline{V}_N$. In particular, for a perturbation of size Δs (of any size, not necessarily infinitesimal), $V_n(s+\Delta s) = V_n(s) + \Delta s$, and hence $\partial \overline{V}_N / \partial s = 1$ is an unbiased estimator for $\partial E\left[\overline{V}_N\right]/\partial s$. Intuitively, the shape of sample paths are unaltered by changes in s if q is held constant; the entire sample path is merely shifted by the size of the change. The IPA estimator can also be obtained by simply differentiating the recursive relationship (7), noting that D_n does not depend on s or q:

$$\frac{dV_{n+1}}{d\theta} = \begin{cases} \frac{dV_n}{d\theta} & \text{if } V_n \geq s, \\ 1 & \text{if } V_n < s. \end{cases}$$

for either $\theta = s$ or $\theta = q$. Taking $V_0 = S = s + q$, the expression reduces to 1 for all n, which is in accord with the sample path analysis.

On the other hand, a change in q with s held fixed may cause a change in the set of ordering decisions, resulting in radical changes in the sample path and hence in the sample performance function $\overline{V}_N$. Thus, SPA is required to derive an unbiased derivative estimator with respect to $\theta = q$. An SPA estimator for $\partial E\left[\overline{V}_N\right]/\partial s$ that can be easily and efficiently estimated from the original sample path is given by

$$1 + \frac{1}{N}\sum_{n \leq N: V_n < s} \frac{f(V_n + D_n - s)}{1 - F(V_n + D_n - s)}\left[s - E[D] - \overline{V}_N\right].$$

Real-World Application Example

In the October 30, 2000 issue of *Fortune* magazine, an article entitled, “New Victories in the Supply-Chain Revolution” (Siekman 2000) describes “a classic distribution challenge: how to avoid lost sales without incurring the cost of carrying extra inventory” when Caterpillar, the “world’s largest builder of construction equipment . . . posed daunting supply chain questions” regarding the distribution of a new line of compact construction machines, specifically related to determining appropriate inventory levels for the U.S. market. “Among the techniques . . . used to attack this complex (supply chain inventory control) problem was . . . infinitesimal perturbation analysis, for which no complete explanation is possible for the faint-hearted or mathematically disadvantaged.”

Historical Notes

PA was developed by Ho et al. (1979) when the first author was consulting on a real-world buffer design problem for a Fiat Motor Company serial production line. The single-server queue example was first considered in Suri and Zazanis (1988), and the inventory example in Fu (1994). The other area in which PA has been most widely used after queueing and inventory is financial engineering, where IPA is called the pathwise method in Glasserman (2004); see also Fu and Hu (1995). Other applications include PERT networks, dams, insurance, preventive maintenance, statistical process control, and traffic light signal control; see Ho and Cao (1991), Fu and Hu (1997), and Fu (2006) for examples and references.

See

- ▶ Inverse Transform Method
- ▶ Score Functions
- ▶ Sensitivity Analysis
- ▶ Simulation of Stochastic Discrete-Event Systems
- ▶ Simulation Optimization
- ▶ Variance Reduction Techniques in Monte Carlo Methods

References

Brémaud, P., & Vázquez-Abad, F. J. (1992). On the pathwise computation of derivatives with respect to the rate of a point process: The phantom RPA method. *Queueing Systems: Theory and Applications, 10*, 249–270.

Cao, X. R. (1994). *Realization probabilities: The dynamics of queueing systems*. Boston: Springer.

Cassandras, C. G., & Larfortune, S. (2008). *Introduction to discrete event systems*. New York: Springer.
Cassandras, C. G., & Strickland, S. G. (1989). On-line sensitivity analysis of Markov chains. *IEEE Transactions on Automatic Control, 34*, 76–86.
Dai, L. Y., & Ho, Y. C. (1995). Structural infinitesimal perturbation analysis for derivative estimation in discrete event dynamic systems. *IEEE Transaction on Automatic Control, 40*, 1154–1166.
Fu, M. C. (1994). Sample path derivatives for (*s, S*) inventory systems. *Operations Research, 42*(2), 351–364.
Fu, M. C. (2006). Gradient estimation. In S. G. Henderson & B. L. Nelson (Eds.), *Handbooks in operations research and management science: Simulation, chapter 19* (pp. 575–616). Amsterdam: Elsevier.
Fu, M. C. (2008). What you should know about simulation and derivatives. *Naval Research Logistics, 55*(8), 723–736.
Fu, M. C., Hong, L. J., & Hu, J. Q. (2009). Conditional Monte Carlo estimation of quantile sensitivities. *Management Science, 55*(12), 2019–2027.
Fu, M. C., & Hu, J. Q. (1995). Sensitivity analysis for Monte Carlo simulation of option pricing. *Probability in the Engineering and Informational Sciences, 9*(3), 417–446.
Fu, M. C., & Hu, J. Q. (1997). *Conditional Monte Carlo: Gradient estimation and optimization applications*. Boston: Kluwer Academic.
Gaivoronski, A., Shi, L. Y., & Sreenivas, R. S. (1992). Augmented infinitesimal perturbation analysis: An alternate explanation. *Discrete Event Dynamic Systems: Theory and Applications, 2*, 121–138.
Glasserman, P. (1991). *Gradient estimation via perturbation analysis*. Boston: Kluwer Academic.
Glasserman, P. (2004). *Monte Carlo methods in financial engineering*. New York: Springer.
Gong, W. B., & Ho, Y. C. (1987). Smoothed perturbation analysis of discrete-event dynamic systems. *IEEE Transactions on Automatic Control, AC-32*, 858–867.
Ho, Y. C., & Cao, X. R. (1991). *Perturbation analysis and discrete event dynamic systems*. Boston: Kluwer Academic.
Ho, Y. C., Cao, X. R., & Cassandras, C. G. (1983). Infinitesimal and finite perturbation analysis for queueing networks. *Automatica, 19*, 439–445.
Ho, Y. C., Eyler, M. A., & Chien, T. T. (1979). A gradient technique for general buffer storage design in a serial production line. *International Journal of Production Research, 17*, 557–580.
Ho, Y. C., & Li, S. (1988). Extensions of infinitesimal perturbation analysis. *IEEE Transactions on Automatic Control, AC-33*, 827–838.
Hong, L. J. (2009). Estimating quantile sensitivities. *Operations Research, 57*(1), 118–130.
Shi, L. Y. (1996). Discontinuous perturbation analysis of discrete event dynamic systems. *IEEE Transactions on Automatic Control, 41*, 1676–1681.
Siekman, P. (2000). New victories in the supply-chain revolution. *Fortune*, (October 30).
Suri, R., & Zazanis, M. A. (1988). Perturbation analysis gives strongly consistent sensitivity estimates for the M/G/1 queue. *Management Science, 34*, 39–64.
Vakili, P. (1991). Using a standard clock technique for efficient simulation. *Operations Research Letters, 10*(8), 445–452.

Perturbation Methods

Procedures that modify the constraints of a linear-programming problem so that all basic feasible solutions will be nondegenerate, thus removing the possibility of cycling in the simplex method. The modification can be either explicitly done by adding small quantities to the right-hand sides or implicitly by using lexicographic procedures.

See

- ▶ Cycling
- ▶ Degeneracy
- ▶ Lexicographic Ordering

Petroleum Refining

David S. Hirshfeld
MathPro Inc., Bethesda, MD, USA

Introduction

By many financial and physical measures, the petroleum industry is the world's largest industry. The industry's operations comprise a global supply chain that produces, transports, refines, and distributes more than 85 million barrels of oil per day – nearly 5 billion tons per year.

Because of its scale, global scope, and huge capital requirements, the petroleum industry is populated with many large, vertically-integrated companies (many of them national oil companies) with global operations. The industry is highly competitive because it has many participants and because it produces basic commodities (e.g., gasoline, diesel fuel, petrochemical feedstocks, etc.) that are difficult to differentiate by brand. The industry's huge volume and low margins mean that even small changes in operating costs have important effects on operating results. The petroleum industry is a leader in the development and application of new technology; it develops and applies advanced technologies in every phase of operations. Consequently, the industry employs large numbers of scientists, engineers, and applied mathematicians, many with advanced degrees.

For these and other reasons, the petroleum industry has been a pioneer in the application of OR/MS across all of its primary operations and has successfully applied virtually every OR/MS tool in these operations. During the 1960s and 1970s, most large integrated oil companies had strong OR/MS groups or departments with concentrations of expertise in linear programming, simulation, and statistical analysis (Baker, 2000). These groups consistently stretched the limits of OR/MS tools and methods, and they provided the impetus and the financial support for many advances in OR/MS software tools and analytical methods. Most of these groups no longer exist. But even so, OR/MS applications in the petroleum industry are ubiquitous and fully embedded in the various business functions that use them. Nowhere is this more evident than in the petroleum refining sector.

OR/MS and Petroleum Refining

Petroleum refining is a unique and critical link in the petroleum supply chain. The other links add value mainly by performing spatial transformations on petroleum (e.g., lifting crude oil to the surface; moving crude oil from oil fields to storage facilities and then to refineries; moving refined products from refinery to terminals and end-use locations, etc.). Refining adds value by performing chemical transformations and blending operations on petroleum – converting crude oil (which in itself has little end-use value) into a broad spectrum of valuable refined products. The primary economic objective in refining is to maximize that added value.

Petroleum refineries are large, continuous-flow process plants with extremely complex processing schemes for processing multiple crude oils and other input streams into a large number of refined (co-) products, most notably LPG, gasoline, jet fuel, diesel fuel, petrochemical feedstocks, home heating oil, fuel oil, and asphalt. Each refinery has a unique configuration and operating characteristics, determined primarily by its location, vintage, preferred crude oil slate, and market requirements for refined products. More than 660 refineries, in 116 countries, are currently in operation; virtually every one has OR/MS tools, including optimization models, embedded in its operations.

Since the earliest days of OR/MS and continuing to the present, refining has been a particularly rewarding domain for applying OR/MS methods in general, and linear programming (LP) and its extensions in particular (mixed integer programming (MIP), special ordered sets (SOS1 and SOS2), and successive linear programming (SLP), etc.).

OR/MS Applications in Petroleum Refining

Baker (2000) reports, “The refining industry began using linear programming (LP) shortly after its invention (Bodington and Baker 1990). In the early 1950s, many major oil companies began using LP-based product blending models (Charnes et al. 1952) which severely tested the available computational capabilities of that time. As computer capabilities expanded, so did the scope of LP models, encompassing whole refineries (Symonds 1955) and the US refining industry (Manne 1958).”

“The nonlinear nature of petroleum and chemical processes was first incorporated by Shell Oil via successive linear programming (SLP), a straightforward technique based on the iterative solution of linearized models (Griffith and Stewart 1961). SLP... was applied by most major companies in the 1960s (Baker and Lasdon 1985). Distributed recursion (DR), a specific form of SLP dealing with the distribution of nonlinear error terms across [multiple] blended pools, is widely used in contemporary models of petroleum refining.”

“Literally… every other form of nonlinear optimization has been applied in the [refining] industry. Lasdon and Waren (1980) provided a comprehensive survey of applications. Production planning and scheduling has seen a wide variety of hybrid approaches combining mathematical programming, expert systems, decision support systems, forecasting techniques and simulation. Klingman et al. (1987) describes the integrated logistics system developed at Citgo. A combination of network flow algorithms, mixed-integer programming, and decision support were applied to ship scheduling at Ethyl Corporation (Miller, 1987). Brown et al. (1987) reports on a vehicle loading and routing system developed for Mobil Oil. The design and development of integrated systems for planning and scheduling is an area of active interest both in academic and industrial settings (Baker 1994)."

Today, mathematical programming and other OR/MS techniques are embedded in numerous refining sector functions, including (in roughly decreasing order of time horizon):

- Capital investment planning
  - Economic evaluation of alternative designs for new refineries
  - Evaluation of alternative configurations for refinery upgrading projects
- Process design
- Tactical planning
  - Evaluation of inter-company product exchanges and processing agreements
  - Optimization of multi-period operations of multi-refinery, multi-terminal logistics systems
  - Evaluation of new processes and technologies
  - Regulatory compliance
- Operations planning
  - Crude oil valuation and supply planning
  - Crude oil cargo selection (Pawde and Singh 2010)
  - Development of quarterly and monthly refinery operating plans
  - Integration of refinery operations and refined product distribution (Guyonnet et al. 2009)
  - Concurrent multi-product blending
- Operations scheduling
  - Process sequencing
  - Inventory (tankage) management
  - Batch blending of refined products
- Process control

The planning and design applications have time horizons measured in years or months, are forecast-driven, and can return solutions in which multiple operations or operating modes employing the same resources or facilities are executed in the given time period. Scheduling applications, on the other hand, have much shorter time horizons (weeks or days), are order- or sequence-driven, and recognize operating policies or physical constraints on the utilization of specific facilities – e.g., only one activity or operation at a time can be performed in a particular facility. Plans returned by planning models may not be physically implementable without being subjected to a detailed scheduling analysis.

Refining organizations use their refining optimization models across many planning horizons:

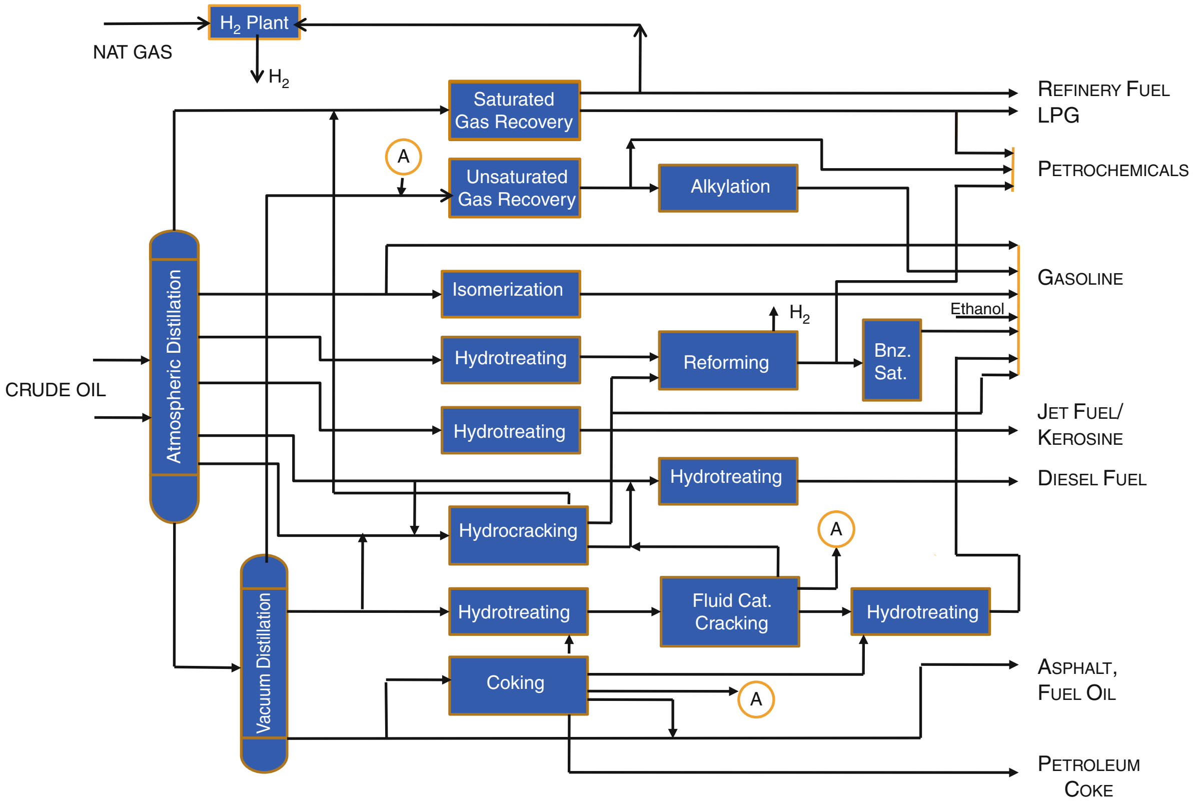
- **Long-term** (3+ years): capital investment planning, regulatory compliance, restructuring
- **Annual**: annual budgeting, evaluation of term contracts for crude supply and product sales, maintenance and turn-around planning
- **Quarterly/monthly**: operations planning to meet product demands and seasonal transitions in product specifications, evaluation of spot transactions for crude purchases and product sales, estimation of dispatches to product pipelines and tankers
- **Weekly**: scheduling operations and batch blending to make optimal use of crudes on hand and available processes

Refinery planning applications are practiced not only by refinery organizations but also by other organizations having interest in the refining sector, such as engineering firms, independent technology providers (e.g., process licensors), catalyst and chemical manufacturers, and consulting firms. Government agencies also apply LP to analyze refining operations, for various purposes – for example, the U.S. Environmental Protection Agency in estimating the costs of new regulatory standards for transportation fuels, and the U.S. Energy Information Administration in producing its annual projections of U.S. energy supply and demand).

Refining Operations and the Driving Forces for Refinery Modeling

Understanding the rationale for and benefits of OR/MS methods in refining industry requires some understanding of refining itself (The National Petroleum Council (2000) Web site includes an excellent tutorial on the fundamentals of refinery operations).

Figure 1 is a highly simplified flow chart of a notional complex refinery, illustrating a typical pattern of oil flow through the refinery – from the crude oil distillation unit that separates crude oil into various boiling range fractions, or cuts, through the various downstream processing units that chemically transform these fractions into blendstocks (the refinery streams that are the constituents of blended products) and ultimately to product blending. For purposes of



Petroleum Refining, Fig. 1 Simplified Flow Chart of a Notional Refinery

this discussion, the importance of Fig. 1 is not in its details, but in the overall picture it conveys of the complexity of refining operations in general.

Several broad aspects of refining operations suggested by Fig. 1 merit comment in the context of refinery modeling applications.

- Refinery operations are extremely complex.

Figure 1 only hints at the actual complexity of refinery operations – with respect to the physical facilities of the refinery, the interaction of these facilities with one another, and the range of operations of which they are capable. The complexity is such that refinery operations can be fully understood only with formal, refinery-wide models and can be optimized, in an economic sense, only through the use of mathematical programming.

Refiners can change the operations of their refineries to respond to the continual changes in crude oil and product markets, but only within physical limits defined by the performance characteristics of their refineries and the properties of the crude oils they process. Mathematical programming models of refinery operations that express these physical constraints are the only reliable means of generating achievable (i.e., feasible) and economic (i.e., optimal) responses to changes in market environment.

- Refineries produce a wide range (or slate) of products – actually co-products.

Refineries produce a range of co-products not only because of market demand for the various products but also because of the constraints imposed by the refining facilities themselves. Refiners need to know the marginal cost of production for each refined product, because these marginal costs are the primary determinants of the products' spot prices – the prices at which products change hands at the refinery gate. Mathematical programming models of refinery operations routinely produce rigorous estimates of marginal production costs that are well grounded in theory, for every co-product produced (The solution values for certain of the dual variables in a refinery

Petroleum Refining, Table 1 Classification of Refining Processes

Primary Classes of Refining Processes in Complex Refineries

Class	Function	Examples
Crude distillation	Separate crude oil charge into boiling range	Atmospheric distillation
	fractions for further processing	Vacuum distillation
Conversion	Break down ("crack") heavy crude fractions into lighter,	Fluid cat cracking
	higher-valued streams for further processing	Coking, Hydrocracking
Upgrading	Enhance the blending properties (e.g., octane) and value	Reforming
	of gasoline and diesel blendstocks	Alkylation, Isomerization
Treating	Remove hetero-atom impurities from refinery streams	Hydrotreating
	and blendstocks	Caustic treating
Separation	Separate, by physical or chemical means, constituents	Fractionation
	of refinery streams for further processing	Extraction
Blending	Combine blendstocks to produce finished products that	
	meet product specifications and environmental standards	
Utilities	Supply refinery fuel, power, steam, oil movements,	Power generation
	storage, emissions control, etc.	Sulfur recovery

model are precisely the marginal values in question). Indeed, mathematical programming is essentially the only practical and useful tool for computing the marginal costs of refined products.

All of this was readily apparent to the engineers and applied mathematicians working in the refining sector in the 1950s and provided the impetus for the early adoption of linear and mathematical programming throughout the refining industry.

Refinery Processes and Operations

Complex, world-class refineries (including virtually all U.S. refineries) comprise as many as fifty or more distinct refining processes, which carry out multiple physical and chemical transformations to convert crude oil into a broad slate of refined products. Despite their number and diversity, refining processes can be thought of in terms of a few broad classes based on their functions, as shown in Table 1.

Crude Oil and the Crude Oil Distillation Process

Crude oil distillation, the process at the front end of every refinery, regardless of size or overall configuration, has a unique function that affects all of the processes downstream of it. In a refinery model, the representation of crude oil properties and of the crude distillation process in a refinery model influences all of the other process representations in the model.

Crude oil comprises tens of thousands of chemical compounds (primarily hydrocarbons). These compounds range from the very light – low molecular weight, simple structure, low density, low boiling point (<60° F) – to the very heavy – high molecular weight, complex structure, high density, high boiling point (>1000° F).

Each of the more than 1,500 crude oils in commerce has its own unique signature, with respect to composition, proportions of light and heavy components, and physical properties. The unique composition and properties of a crude oil largely determine its value as a refinery input and the range of refined products that a given refinery can produce from it.

The crude distillation unit in a refinery accepts a combination of different crude oils and separates it into a number of streams (known as crude fractions or cuts). Each fraction leaving the crude distillation unit (1) is defined by a unique boiling point range (e.g., 180°–250° F, 250°–350° F, etc.), (2) contains material from each crude oil fed to the crude distillation unit, and (3) is made up of hundreds of distinct hydrocarbon compounds, all of which have boiling points within the cut range. An essential simplifying assumption in the analysis of refining operations is that the crude distillation unit makes "sharp" cuts – that is, any given hydrocarbon species in the crude oil mixture is present in one and only one cut (i.e., there is no "overlap" between the crude fractions leaving the crude distillation unit).

Each crude fraction leaving the crude distillation goes to a different refinery process for further processing (Fig. 1). The highest boiling fractions of the crude, collectively known as the heavy ends, have relatively little economic value – indeed lower value than the crude oil from which they come. Refineries must convert, or upgrade, these heavy ends into more valuable light products (gasoline, jet fuel, diesel fuel, etc.).

Stream Properties and Refining Processes

In a refinery model, the specification of the temperature ranges of the cuts and the representation of the various properties of the crude fractions exerts a strong influence on the representations of all of refining processes downstream of the crude distillation and on the results returned by the model.

In general, each refining process handles multiple feed streams and produces multiple outputs (co-products). The yields of the co-products, their physical and chemical properties, and the direct operating costs of each process depend on the properties of the input streams (which in turn depend on the mixture of crude oils processed and the temperature ranges of the crude cuts). Consequently, analyzing refinery operations requires keeping track of not only the various streams flowing through the refinery but also numerous properties associated with each stream.

Tracking stream properties is essential in analyzing the blending operations at the back end of every refinery. Refineries produce a diverse set of co-products (e.g., gasolines, jet fuel, diesel fuels, petrochemical feedstocks, etc.); large, complex refineries may produce as many as forty distinct products. Most of these products are blends of various streams produced in crude distillation or in the downstream processes (usually five to ten refinery streams per product). Each product is blended to meet a vector of specifications on the products' properties (e.g., density, sulfur content) and performance characteristics (e.g., octane, emissions from vehicle tailpipes, etc.). These specifications represent industry standards and government regulations.

The Content of Refinery LP/MP Models

Structure

An LP or MP model of a single refinery in a single time period is essentially an assembly of

- Equations and inequalities representing
  - Volume balances on refinery inputs, refinery-produced streams, and refinery outputs (volume supplied + volume produced = volume consumed + volume blended or sold)
  - Mass balances and energy balances (conservation of mass and energy)
  - Blending property balances linking individual refinery streams and their blending properties to specification-blended product pools
  - Accounting identities to capture refinery-wide operating costs, consumption of energy and utilities, and generation of effluents (including CO_2)
  - Upper limits on the through-put capacity of the various refining processes
  - Special constraints reflecting internal technical restrictions or limitations
  - Special constraints reflecting external requirements
  - Regulatory standards (such as the federal and California standards for reformulated gasoline).
- Variables representing
  - Volumes of refinery inputs, such as crude oil purchases
  - Volumes of refinery streams flowing into or out of each process unit (such as those shown in Fig. 1) at specified operating conditions
  - Volumes of produced refinery streams going to each blended product pools
  - Volumes of finished products leaving the refinery
  - Amounts of new refinery process capacity (if any) added through capital investment

Multi-time-period models contain, in addition to the above elements, equations and variables representing inventory transfers from one time to the next of crude oils, other refinery inputs, certain intermediate refinery streams, and finished products.

Multi-refinery models contain, in addition to the above elements, equations and variables representing the transport of refined products from the refineries to individual destinations (product terminals, end-use sites, etc.) or destination regions, through various capacitated transportation modes.

In all of these variants, the objective function usually represents gross profit or, as it sometimes called, profit contribution:

Refinery netback minus (the sum of direct operating costs + capital recovery charges)

where

- **Refinery netback** is the net revenues (price*quantity) received by the refinery from the sale of all refined products
- **Direct operating costs** include the total purchase costs (price*quantity) of crude oil and other refinery inputs, purchased utilities, and catalyst and chemicals consumption; inventory carrying costs (in multi-period models); transportation costs for product movements to demand sites (in multi-refinery models), and regulatory compliance costs
- **Capital recovery charges** denote return on un-depreciated refinery investment, per unit of throughput.

In multi-period models, the profit contribution terms for future time periods can be discounted by multiplying them by a discount rate factor: (1+ discount rate)$^{-t}$, where t is the time-period index.

Models of refinery operations contain distinct representations of each of the refining processes that have a significant effect on the refinery's economics. A complex refinery can comprise forty or more such processes. Each process (or process/refinery combination, in a multi-refinery model) is represented in a discrete sub-matrix of the overall model. Each process sub-matrix consists of one or more operating mode or input/output variables, any number of which can be active in a given solution. Each operating mode variable intersects certain equations representing volume balances on the streams flowing into and out of the process, energy balances, and accounting relationships. The vector of input/output coefficients associated with each operating mode variable denote the quantities of individual inputs (refinery streams, utilities, capacity, costs) and outputs (different refinery streams) per unit of process throughput in a particular operating mode, as well as the relevant properties of the output streams.

Depending on the number of processes and refinery streams represented, a typical single-refinery, single-time-period LP model contains about 1,500–5,000 constraints, and 5,000–15,000 variables. Refinery models have highly structured matrices, composed of the various process and blending sub-matrices, linked by the volume balance and property balance constraints. The matrices are relatively dense, but have low super-sparsity (because the input/output coefficients in the process representations tend to be unique).

Coefficients

The coefficients for the crude oil distillation sub-matrix usually are drawn from *crude oil assays*. A crude oil assay is an assembly of data on the composition and property of a whole crude oil and of 15–20 boiling range fractions of that of that crude, developed through laboratory testing.

Crude assays exist for all crude oils in commerce; many, but not all, of these assays are in the public domain.

Commercial software products called crude oil assay managers with associated assay libraries are widely used to generate the coefficients for representing the crude oil distillation process in a refinery model, with user-specified boiling ranges for the crude fractions.

The coefficients for the sub-matrices representing the refining processes are refinery-specific in most models and are derived, directly or indirectly, from experimental data. Depending on the process, the data may come from laboratory testing, pilot plant operations, refinery-level plant testing, refinery accounting systems, and process simulators (detailed engineering models of individual refining processes). In general, all of these sources of refinery data are proprietary.

Some non-proprietary, generalized correlations and data for characterizing refining processes are available in the open literature, primarily in a few textbooks (e.g., Maples (2000), Gary et al. (2000)) and articles in refining industry trade journals.

Populating a refinery optimization model with realistic input/output coefficients is a highly specialized undertaking, requiring considerable knowledge of refinery operations and refining technology – subjects that are at some remove from operations research.

Nonlinearities in Refinery Models

To this point, this overview of refinery optimization models seems to imply that refining operations are linear in nature and therefore can be suitably represented as linear programming models. Refining operations are subject to mass balance, energy balance, and volume balance constraints, all of which are linear, as are the constraints that govern multi-ingredient blending to meet product specifications (as long as the blending is simply physical mixing with no chemical interactions between ingredients). Consequently, refinery optimization was a natural pioneering application for linear programming. And even today, LP remains the optimization method of choice for many refinery modeling applications.

However, refinery operations actually embody many nonlinear phenomena, some of which can have a strong influence on refining operations and economics. Almost from the beginning, a steadily increasing number of refining organizations have sought to enhance their capabilities to capture these nonlinearities in their refinery optimization models and thereby more accurately represent the true capabilities and limitations of their refining facilities.

Some of the nonlinearities of interest are economic in nature and bear on the objective function; others involve underlying physical processes and relationships and bear on the constraint set. Many of these nonlinearities, including the five discussed below, are incorporated readily in refinery models, facilitated in many instances by the capabilities of commercial solvers.

Investments in New Refining Capacity

Existing refineries often invest in additional processing capacity – either new process units or expansion of existing ones – in order to increase total production capacity, produce new products, upgrade the value of existing products, or comply with new regulatory standards bearing on product quality or performance characteristics.

Often, the capacity added for a given process is represented by a continuous variable (whose value is expressed in a capacity measure, such as K barrels/day), and the corresponding investment is approximated by multiplying this variable by a constant investment rate coefficient (whose value is in $/(barrel/day)).

$$\mathbf{I} = \mathbf{a}^{*}\mathbf{Q} \qquad (1)$$

where **I** is the investment (in K$), **Q** is the capacity added (in $/barrel/day), and **a** is the investment rate factor ($/(barrel/day)). The value of the investment rate factor depends on the refining process and the refinery's location.

However, the capital investment required to add new refining capacity enjoys economies of scale; that is, the investment per unit of added capacity is not a constant, but decreases with increasing total amount of added capacity. The standard relationship between the amount of new capacity added and the required capital investment is

$$\mathbf{I} = \mathbf{b}^{*}\mathbf{Q}^{\beta} \qquad (2)$$

where **I** is the investment (in K $), **Q** is the capacity added (in K barrels/day), **b** is a constant whose value depends on the refinery's location, and β is an exponent whose value depends on the refining process in question. Most refining processes have a β value in the range of 0.6–0.7.

Equation (2) is a non-convex function. It can be represented in a refinery MP model in one of several ways.

One approach is to (1) assign a set of binary (0–1) variables to each of three or four standard levels of new capacity addition (e.g., 10 K barrels/day, 20 K barrels/day, etc.) for each refining process that is a candidate for investment and (2) for each such set, add a constraint specifying that at most one of the variables in the set can take on the value 1 in an optimal solution (or, equivalently, define the set of binary variables for each refining process as a Special Ordered Set Type 1 (SOS1)). Each of the binary variables carries a coefficient denoting the capital investment for the capacity addition it represents, obtained from the (2) for each process.

Another approach is to represent (2) for each process that is a candidate for investment as a piecewise linear function by means of a Special Ordered Set Type 2 (SOS2) for each such process.

Semi-Continuous Quantities

In many situations, restrictions exist on the minimum and maximum volume of a particular flow or the minimum and maximum extents to which a particular operation can be performed. For example, pipeline off-takes from a refinery are subject to the pipeline's regulations on the minimum and maximum size shipments that it will accept. Similarly, purchases of tanker-borne crude oil are subject to volume to volume limits determined by the size of the tanker and its cargo compartments.

These and similar constraints can be represented in refinery models by means of semi-continuous variables: variables that can be either zero or continuous within a range defined by a strictly non-zero lower bound and (optionally) an upper bound. Semi-continuous variable capability is available in most commercial having mixed-integer-programming (MIP) capability.

Quality Blending

In the canonical product blending problem, the ingredients blend linearly with respect to the blend properties that are subject to limits (specifications). That is, the properties of the blended product requirements are the weighted averages of the corresponding properties of the various ingredients. This is linear blending.

Many refinery models represent the blending of refined products to specifications just that way. However, there is more to the specification blending of refined products than simple linear blending. Some of the specifications to which refined products are blended pertain to purely physical properties (e.g., sulfur content, density); other to chemical properties (e.g., octane, volatility, etc.). Blending to specifications on physical properties is indeed linear, as defined above. However, blending of chemical properties often is not linear, because of the interactions among different chemical interactions that occur when individual ingredients (*blendstocks* in refining parlance) are blended together. For example, consider two gasoline blendstocks, one having 90 octane, the other 70 octane. A 50/50 blend of the two might yield a blend octane of, say, 82 or 77 (not 80), depending on the chemical interactions involved. Moreover, the blend octane may vary with the relative amounts of the two blendstocks. This is nonlinear blending.

Several techniques are available for representing nonlinear blending. The most widely used one involves the use of blending indices in place of blendstock properties. A blending index for a given nonlinear property is an empirically determined function of that property such that the function blends linearly, even though the property itself does not. For example, consider the property Reid Vapor Pressure (RVP), a standard measure of gasoline volatility. RVP blends nonlinearly, but the RVP Index, defined here, blends linearly.

$$\text{RVP Index} = \text{RVP}^{\rho} \qquad (3)$$

where the value of the exponent ρ is about 1.17 (Different refiners may use slightly different values for ρ).

Some blending indices involve more complicated functions of the underlying property. For example, Pour Point (PP), a measure of diesel fuel's ability to flow at low temperature, has a Pour Point Index given by:

$$\text{PP Index} = \text{EXP}[1.85 + 0.042^{*}(\text{PP})] \qquad (4)$$

Many gasoline and diesel fuel blending properties are represented by such blending indices in refinery models.

Gasoline octane blending is a special instance of nonlinear blending for two reasons. First, octane has a relatively high marginal refining cost; refiners do not wish to "give away" octane in the course of meeting the octane standards. Second, the blending octane of a gasoline blendstock (i.e., the apparent octane contribution of the blendstock to the finished blend) is a function not only of the blendstock's native octane but also the composition of the finished blend. The refining industry has developed special methods, based on laboratory data, to estimate blend octanes over a range of compositions. These methods, outlined by Maples (2000), are beyond the scope of this article.

Pooling

Pooling is the mixing or commingling of multiple streams (crude fractions or refinery streams) into a new stream (the pool), whose properties (e.g., density, sulfur content, etc.) are the volume-weighted averages of the properties of the individual streams entering the pool:

$$\mathbf{Q}_j V = \Sigma_i q_{ij} V_i \quad \Rightarrow \quad \mathbf{Q}_j = \Sigma_i q_{ij} V_i \;/\; \Sigma_i V_i \qquad (5)$$

where **V** is the volume of the pool stream, $\mathbf{Q}_j$ is the j^{th} property (e.g., density) of the pooled stream, V_i is the volume of the i^{th} stream making up the pool, and q_{ij} is the j^{th} property of that stream.

The q_{ij} are constant coefficients, but **V** and the V_i are variables, whose values are known only when the model returns a solution. Thus, the properties (**Q_j**) of the pooled stream are nonlinear function of model variables and can be determined only after a solution is in hand.

Consequently, it is not possible to define exact representations of those effects on downstream refining operations that depend on the properties of the pooled stream. These effects reside in the refining process and specification blending sub-matrices. Thus, not only do the optimal volumes of the pooled stream, **V**, and the streams making up the pool, V_i, depend on the properties of the pool stream, but also the economic value of the pool stream **V**.

The original, or traditional, approach to formulating refinery models does not address pooling at all – not because the problem was not recognized but because the analytical tools needed to address it were not then at hand. In the traditional approach (still widely used), crude distillation and each of the downstream refining processes are represented in discrete sub-matrices. In the crude distillation sub-matrix, each crude oil is represented by its own input/output vector, in which the output coefficients are the volumetric yields of the various cuts. This representation implies that (1) the various crude oils, each with their own properties and yield patterns, are segregated from one another as they go through the crude distillation unit and (2) the boiling range cuts from the various crude oils are likewise segregated from one another as they move to the downstream processes. In the downstream process sub-matrices, each feed is attributable to a particular crude oil and each is represented by its own input/out vector. This scheme represents each process operating as if it were processing a group of segregated feed streams, each with its own operating mode, rather than one pool stream.

Refinery models formulated in this way tend to contain many more stream flow variables, and many more blendstock variables and blending options, than there are in the “real” refinery. This can lead, in certain situations, to over-optimization – the model’s returning solutions indicating better refining economics than the real refinery can achieve.

Explicit representation of the stream pooling that occurs in real refineries calls for special model formulation and solution techniques. The most widely used modeling technique is called Distributive Recursion (DR), a variant of SLP developed expressly to deal with the pooling problem in models of refining and other process flow industries. First developed in the late 1970s, DR has come into increasingly wide use as the required software tools have become more widely available.

In DR, the model user provides initial estimates of the Q_i for all of the pool streams. The procedure uses these estimates to conduct an initial solution pass, which returns (1) the downstream dispositions and marginal value of each pool and (2) the volumes, V_i, of each stream entering each pool. Using the new set of V_i values, the DR procedure re-estimates the various pool qualities. The difference between the n^{th} and $n + 1^{st}$ estimates for a given pool is called its quality error. DR distributes each quality error across the various downstream dispositions of each pool and initiates a new solution pass incorporating the new estimates of pool qualities and quality errors. DR conducts a series of such solution passes that seek to converge to an optimal solution in which the quality errors are driven to zero (to within a user-specified tolerance).

Performance of Refining Processes

In the original, or traditional, approach to formulating refinery models, each downstream process is represented in a discrete sub-matrix. Each process sub-matrix comprises a set of variables (vectors), each denoting a unique combination of (1) a segregated (not pooled) feed stream to the process and (2) a particular operating mode for the process (defined by physical operating conditions, such as temperature). Each such variable has a unique set of input/out coefficients, defining the operation of the process. This representation implies that (1) processes behave linearly, independent of the composition and properties of their feeds and (2) each (notionally) segregated stream can be processed at its own set of operating conditions as it flows through the process. In reality, process performance depends on the properties of the pooled feed to the process.

With the advent of DR, some refining companies sought a more rigorous representation of refining processes that used pooled input streams and captured the effects of input stream properties on the yields and properties of the output streams. This effort led to the *base-delta* (B-D) approach to representing refining processes in optimization models (Bodington 1995).

In the B-D approach, each downstream process is represented in a discrete sub-matrix, comprising:

- One or more base vectors, each denoting operation of the process with a typical, or base, feed and a standard, or base, operating mode.

  The input coefficients on the base vector(s) denote those properties of the pooled feed that affect the yields and properties of the process outputs. The output coefficients denote the yields and properties of the various outputs, when the process is operating at base conditions.
- A set of delta vectors for each base vector. The solution values taken on by the various delta vectors are determined as part of the overall model solution obtained via DR.

  The coefficients on each delta vector denote the effects of a small change in one pooled feed property (relative to the base property) on the yields and properties of the various outputs. Each delta vector coefficient is, in effect, the partial first derivative of a particular process output property with respect to an input property. The set of all delta vector coefficients for a process is equivalent to a Jacobian matrix for the process.

Both the base yield coefficients and the delta vector coefficients usually are generated by means of detailed engineering models (called process simulators) of the various processes, or (less likely) by generalized correlations or plant testing. Modern refinery modeling systems that offer DR now provide interfaces to process simulators. These interfaces link the process simulators directly to a refinery optimization model and allow them to be invoked at each DR solution pass to dynamically update the some or all of the delta coefficients in response to the current DR solution. Use of this facility increases the likelihood of reaching a local optimum.

Finally, the traditional representation of crude distillation, the refinery's front-end process, treats the cut point temperatures of the various crude fractions (e.g., 160°–250° F for a light naphtha stream) as constants. The advent of DR allows the cut points themselves to be recursed variables, an option that is now widely used.

Comments on Distributed Recursion (DR)

As with any non-linear technique, DR can – and often does – return solutions that are only locally optimal. In particular, the DR procedure requires initial estimates of the properties of each pool and the fractional distributions of each pool to its various downstream dispositions. The specific values of these estimates determine whether the DR procedure converges to a global optimum or to a local optimum. The more pooled streams and the greater the number of pool dispositions in the model, the more likely that the model will return a local optimum.

Capturing the analytical benefits of DR requires considerable software and intellectual resources, including:

- Some means – whether process simulators or sets of correlations – of dynamically representing the effects of process input properties on process output yields and properties;
- An array of special software, including a crude oil assay manager, process simulators (or their functional equivalent), and facilities to execute and control the recursive solution process; and
- Sound model formulation practices, careful estimation of the initial values of stream properties and distributions, and proper settings for the DR procedure's control parameters and tolerances.

Only analysts with access to the necessary system resources and with extensive experience in refinery modeling in general and DR in particular are likely to obtain useful and timely results with DR.

However, many refinery modeling applications do not require the degree of precision that DR is intended to provide in representing the capabilities and limitations of refining facilities. In particular, high accuracy in representing refining facilities may not be warranted in applications, such as tactical and strategic planning, that have planning horizons measured in years, rather than months or weeks. Long planning horizons involve substantial uncertainty regarding crude oil prices, product demands, and other economic factors. These applications place a premium on the ability to rapidly analyze and compare many different model instances, each representing a future economic scenario – as opposed to analyzing a few model instances with greater precision in the representation of refining facilities.

In these situations, the conventional refinery modeling approach, not the DR approach, is usually the method of choice.

Model Management for Refinery Models

Model management is “the care and feeding” of large scale modeling applications. It is a complex of information processing functions that includes model formulation (in an electronic format), data up-dating, case management, matrix generation, optimizer control, solution reporting, model and solution analysis, and model maintenance. All operational use of large scale optimization models involves the performance of these and other functions – whether manually, with an ad hoc collection of software tools, or with a purpose-built software system.

As the size and scope of refinery optimization models increased, the burdens of model management became apparent. The first software tools designed specifically to address elements of model management (as opposed to model solution) were the matrix generation languages fielded in the late 1950s and early 1960s (e.g., Haverly System’s MaGen™ and (later) OMNI™; Bonner & Moore’s MARVEL™ and (later) GAMMA™). These were procedural programming languages with special functionality and features for generating refinery LP models in optimizer input format and generating output reports on model solutions. The matrix generation languages were a large step forward, but they did not provide a full range of model management functionality.

Somewhat later, a new set of software tools for matrix generation and reporting entered commercial use: the algebraic modeling languages: (e.g., GAMS™, AMPL™, MPL™, MODELER™, and AIMMS™). These are symbolic modeling languages, in which the model formulator expresses the model’s constraints and variables symbolically, in an algebra-like syntax. They also provide facilities for model up-dating and report generation.

Starting in the 1950s, many of the major refining companies undertook development of their own comprehensive refinery modeling systems, some using commercial matrix generation languages, others using standard programming languages of the times. Beale (1978) describes British Petroleum’s approach to model management. Palmer et al. (1984) describes the conceptual and design foundations for Exxon’s PLATOFORM™ model management system. At one time, PLATOFORM routinely handled more than one hundred mathematical programming applications in Exxon. Bodington and Baker (1990) reference other companies’ efforts in model management system development.

As a consequence of the waves of consolidation and down-sizing that swept through the petroleum industry starting in the 1980s, most refining companies curtailed or abandoned their efforts to develop and maintain their own model management systems. A few companies still maintain their in-house model management systems. But most refining companies have now supplanted their in-house systems with one of the generalized refinery modeling systems brought into commerce by independent developers (e.g., PIMS™ (AspenTech), GRTMPS™ (Haverly Systems), and RPMS™ (Honeywell Hi-Spec Solutions)).

Commercially available modeling systems must be instantiated with data specific to the refinery of interest: crude oil assays, process capacities and performance characteristics, stream properties, and product specifications. Once instantiated, the generalized refinery modeling systems offer extensive functionality for refinery modeling, including DR (as an option), comprehensive model management functionality, and compatibility with crude oil assay managers, process simulators, spreadsheets, relational databases, and a number of standard commercial solvers.

Concluding Remarks

The petroleum industry pioneered the application of OR/MS across all of its primary operations, and has provided the impetus and the financial support for many advances in OR/MS software tools and analytical methods. This symbiotic relationship is particularly strong in the petroleum refining sector. Since the earliest days of OR/MS, refining has been a particularly rewarding domain for applying OR/MS methods in general, and especially linear programming (LP) and its extensions (in particular, mixed integer programming (MIP), special ordered sets

(SOS1 and SOS2), and successive linear programming (SLP)). As a result, OR/MS applications – especially linear and mathematical programming applications – are ubiquitous and fully embedded in refining operations.

Although petroleum refining is a mature area of application for OR/MS, the tools and methods available to refining industry practitioners continue to improve in terms of speed and functionality. Further advances are likely to come in the realm of model management.

Development and application of optimization models in the refining sector requires deep knowledge of refining technology and economics. Knowledge of optimization algorithms and software tools is necessary but not sufficient for successful application of OR/MS in the refining sector.

See

- ▶ Linear Programming
- ▶ Mathematical Programming
- ▶ Model Management
- ▶ Nonlinear Programming
- ▶ Special-Ordered Sets (SOS)

References

Baker, T. E. (2000). Petrochemical industry. *Encyclopedia of operations research and management science* (2nd ed). Kluwer Academic Publishers.

Baker, T. E. (1994). An integrated approach to planning and scheduling. In D. W. T. Rippin (Ed.), *Foundations of computer-aided process operations* (pp. 237–251). Texas: Austin. CACHE.

Baker, T. E., & Lasdon, L. S. (1985). Successive linear programming at Exxon. *Management Science, 31*, 264–274.

Bammi, D. (1990). Northern border pipeline logistics simulation. *Interfaces, 20*(3), 1–13.

Beale, E. M. L. (1978). Nonlinear programming using a general mathematical programming system. In H. J. Greenberg (Ed.), *Design and implementation of optimization software* (pp. 259–279). The Netherlands: Sijthoff and Noordhoff.

Bodington, C. E. (1995). *Planning, scheduling and control integration in the process industries*. New York: McGraw-Hill.

Bodington, C. E., & Baker, T. E. (1990). A history of mathematical programming in the petroleum industry. *Interfaces, 20*(3), 117–127.

Brown, G. G., et al. (1987). Real-time, wide area dispatch of Mobil tank trucks. *Interfaces, 17*(1), 107–120.

Charnes, A., Cooper, W. W., & Mellon, B. (1952). Blending aviation gasoline–a study in programming interdependent activities in an integrated oil company. *Econometrica, 20*(2), 135–139.

Council, N. P. (2000). *U.S. Petroleum refining: Assuring the adequacy and affordability of cleaner fuels*. Washington, DC: National Petroleum Council.

Edgar, T. F., & Himmelblau, D. M. (1988). *Optimization of chemical processes*. New York: McGraw-Hill.

Findlay, P. L., et al. (1989). Optimization of the daily production rates for an offshore oilfield. *Journal of Operational Research Society, 40*, 1079–1088.

Gary, J. H., Handwerk, G. E., & Kaiser, M. J. (2007). *Petroleum refining technology and economics*. Boca Raton, FL: CRC Press.

Griffith, R. E., & Stewart, R. A. (1961). A nonlinear programming technique for the optimization of continuous processing systems. *Management Science, 7*, 379–392.

Guyonnet, P., Grant, F. H., & Bagajewicz, M. J. (2009). Integrated model for refinery planning, oil procuring, and product distribution. *Industrial and Engineering Chemistry Research, 48*(463–482), 2009.

Hansen, P., et al. (1992). Location and sizing of off-shore platforms for oil exploration. *European Journal of Operational Research, 58*(2), 202–214.

Higgins, J. G. (1993). Planning for risk and uncertainty in oil exploration. *Long Range Planning, 26*(1), 111–122.

Klingman, D., et al. (1987). The successful deployment of management science throughout citgo petroleum corporation. *Interfaces, 17*(1), 4–25.

Lasdon, L. S., & Waren, A. D. (1980). A survey of nonlinear programming applications. *Operations Research, 28*, 102–1073.

Main, R. A. (1993). Large recursion models: Practical aspects of recursion techniques. In T. A. Ciriani & R. C. Leachman (Eds.), *Optimization in industry*. New York: Wiley.

Manne, A. (1958). A linear programming model of the US petroleum refining industry. *Econometrica, 26*(1), 67–106.

Maples, R. E. (2000). *Petroleum refinery process economics* (2nd ed.). Tulsa, Oklahoma: PennWell Corporation.

Miller, D., et al. (1994). A modular system for scheduling chemical plant production. In D. W. T. Rippin (Ed.), *Foundations of computer-aided process operations* (pp. 355–372). Texas: Austin. CACHE.

Miller, D. (1987). An interactive, computer-aided ship scheduling system. *European Journal Operational Research, 32*(3), 363–379.

Palacios-Gomez, F., Lasdon, L., & Enquist, M. (1982). Nonlinear optimization by successive linear programming. *Management Science, 28*(10), 1106–1120.

Palmer, K. H., et al. (1984). *A model-management framework for mathematical programming*. New York: Wiley.

Pawde, M. D., & Singh, S. (2010). “Crude oil cargo selection and time frame of LP optimization,” Petroleum Technology Quarterly, Third Quarter, 2010.

Symonds, G. H. (1955). *Linear programming–the solution of refinery problems*. New York: Esso Standard Oil Company.

Tucker, M. A. (2001). “LP modeling – past, present, and future,” National Petrochemical and Refiners Association (NPRA) 2001 Computer conference, Paper CC-01-153.

PFI

▶ Product Form of the Inverse (PFI)

Phase I Procedure

That part of the simplex method directed towards finding a first basic feasible solution.

See

▶ Artificial Variables
▶ Linear Programming
▶ Phase II Procedure
▶ Simplex Method (Algorithm)

Phase II Procedure

The part of the simplex algorithm that finds an optimal basic feasible solution, starting with Phase I basic feasible solution or an initial basic feasible solution.

See

▶ Linear Programming
▶ Phase I Procedure
▶ Simplex Method (Algorithm)

Phase-type Distribution

▶ Phase-type Probability Distributions

Phase-type Probability Distributions

Marcel F. Neuts
The University of Arizona, Tucson, AZ, USA

The probability distributions of phase-type, or *PH*-distributions, form a useful general class for the representation of nonnegative random variables. A comprehensive discussion of their basic properties is given in Neuts (1981). There are parallel definitions and properties of discrete and continuous *PH*-distributions, but the discussion here emphasizes the continuous case.

The simplest example is the Erlang random variable, which can be expressed as the sum of independent exponentially distributed random variables. As a result, one can construct a realization of an Erlang random variable by going through a series of phases, one for each exponential random variable; hence, the Erlang distribution is a phase-type distribution. Generalizing this phase-type idea governs the movement through the phases by a Markov chain that permits movement back and forth between the interior phases, with the final stage being an absorbing barrier.

More specifically, a probability distribution $F(\cdot)$ on $[0, \infty)$ is of phase type if it can arise as the absorption time distribution of an $(m + 1)$-state Markov chain with m transient states $1, \ldots, m$ and an absorbing state 0. The generator $\boldsymbol{Q}$ of such a Markov chain is written as

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{T} & \boldsymbol{T}^0 \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

where $\boldsymbol{T}$ is a nonsingular $m \times m$ matrix with negative diagonal elements and nonnegative off-diagonal elements. If $\boldsymbol{e}$ denotes a column vector with all components equal to one, then the vector $\boldsymbol{T}^0$ satisfies $\boldsymbol{T}^0 = -\boldsymbol{Te}$. The initial probability vector of the Markov chain is specified as (α, α_0). Without loss of generality, it may be assumed that the generator, $\boldsymbol{Q}^* = \boldsymbol{T} + (1 - \alpha_0)^{-1}\boldsymbol{T}^0\alpha$, is irreducible.

The general formula for the *PH*-distribution $F(\cdot)$ is then

$$F(x) = 1 - \alpha \ \exp(\boldsymbol{T}x)\boldsymbol{e}, \qquad \text{for } \boldsymbol{x} \geq \boldsymbol{0}.$$

The pair $(\alpha, \boldsymbol{T})$ is called a representation of $F(\cdot)$. The *PH*-distribution $F(\cdot)$ has a point mass α_0 at 0 and a density $F'(x) = -\exp(\boldsymbol{T}x)\boldsymbol{Te} = \alpha \exp(\boldsymbol{Tx})\boldsymbol{T}^0$, on $(0, \infty)$. The Laplace-Stieltjes transform $f(s)$ of $F(\cdot)$ is

$$f(s) = \alpha_{m+1} + \boldsymbol{\alpha}(s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{T}^0, \quad \text{for Re s} \geq 0.$$

Its moments λ_ν, $\nu \geq 1$, are all finite and given by $\lambda_\nu = (-1)^\nu \nu!\,\alpha T^{-\nu}\,\boldsymbol{e}$. Some special classes of

PH-distributions are the hyperexponential distributions

$$F(x) = \sum_{\mathrm{v}=1}^{m} \alpha_{\mathrm{v}} \left(1 - e^{-\lambda_{\mathrm{v}} x}\right),$$

which may be represented by $\alpha = (\alpha_1, \ldots, \alpha_m)$, $\alpha_{m+1} = 0$, and $\boldsymbol{T} = -\mathrm{diag}(\lambda_1, \ldots, \lambda_m)$, and the (mixed) Erlang distributions

$$F(x) = \sum_{\mathrm{v}=1}^{m} p_{\mathrm{v}} E_{\mathrm{v}}(\lambda; x),$$

which are represented by $\alpha = (p_m, p_{m-1}, \ldots, p_1), \alpha_{m+1} = 0$, and

$$\boldsymbol{T} = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\lambda & \lambda & \cdots & 0 & 0 & 0 \\ & \cdots & & & & \cdots & \\ 0 & 0 & 0 & \cdots & 0 & -\lambda & \lambda \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\lambda \end{bmatrix}$$

Uses of Phase-type Distributions

The utility of *PH*-distributions is due to their closure properties, which allow standard operations such as convolution and mixing to be represented by matrix operations. Many classical simplifying properties of the exponential distribution have analogs in the matrix formalism for *PH*-distributions. In the analysis of probability models, *PH*-distributions often lead to tractable results without the severe restriction of exponential assumptions. Integrals involving *PH*-distributions also can usually be evaluated by stable recurrence relations or differential equations. Moreover, the phase-type distributions form a dense subset of the probability distributions on $[0, \infty)$, in that any such distribution can in principle be uniformly approximated by a sequence of *PH*-distributions.

Examples of closure properties are:

(a) If $F(\cdot)$ is a *PH*-distribution with representation $(\alpha, \boldsymbol{T})$ and mean $\lambda_1^{'}$, the corresponding delay distribution $F^{*}(\cdot)$ with density $(\lambda_1^{'})^{-1}[1 - F(x)]$ is *PH* with representation $(\boldsymbol{\pi}, \boldsymbol{T})$ where $\boldsymbol{\pi} = (\lambda_1^{'})^{-1} \alpha (-\boldsymbol{T})^{-1}$.

(b) If $F(\cdot)$ (with $\alpha_0 = 0$) is the service time distribution of a stable M/G/1 queue with arrival rate θ and service time distribution $H(\cdot)$ of mean $\mu_1^{'}$, such that $\rho = \theta \mu_1^{'} < 1$, the (steady-state) distribution $W(\cdot)$ of the waiting time is *PH*. Its representation is given by $(\boldsymbol{\gamma}, \boldsymbol{L})$, where $\boldsymbol{\gamma} = \rho \boldsymbol{\pi}$, $\boldsymbol{L} = \boldsymbol{T} + \rho \boldsymbol{T}^0 \boldsymbol{\pi}$. For the *M/PH/1* queue, the distribution $W(\cdot)$ may therefore be computed by integrating a system of linear differential equations, rather than by solving the Pollaczek-Khinchin integral equation.

The fact that any probability distribution on $[0, \infty)$ can be approximated by *PH*-distributions is of somewhat limited practical application, although very good *PH*-approximations to classes such as the Weibull distributions have been obtained. Because of the following general result, that denseness property is, however, of considerable theoretical utility.

Suppose that a stochastic model involves one or more general probability distributions $F_j(\cdot)$, $1 \leq j \leq N$, on $[0, \infty)$, requiring evaluation of a continuous functional $\Phi[F_1(\cdot), \ldots, F_N(\cdot)]$. If an expression for $\Phi(\cdot)$ can be found for the case where $F_1(\cdot), \ldots, F_N(\cdot)$ are *PH*-distributions and if that expression does not explicitly depend on the formalism of *PH*-distributions, then it is also valid for arbitrary distributions $F_1(\cdot), \ldots, F_N(\cdot)$. This result has been used to establish various moment and other formulas in the theory of queues.

There is an extensive literature on phase-type distributions and their applications, including topics such as the structural geometric properties of families of *PH*-distributions, the approximation of other families of distributions by those of phase-type, and the fitting of *PH*-distributions to data. An important characterization of *PH*-distributions was proved in O'Cinneide (1990). Procedures for the approximation by *PH*-distributions are discussed in Asmussen et al. (1992), Johnson (1993) and Schmickler (1992). The appearance of phase-type distributions in some unexpected places in queueing theory was noted in Asmussen (1992).

See

- ▶ Erlang Distribution
- ▶ Hyperexponential Distribution

- ▶ Markov Chains
- ▶ Markov Processes
- ▶ Queueing Theory

References

Asmussen, S. (1992). Phase-type representations in random walk and queueing problems. *Annals of Probability, 20*, 772–789.

Asmussen, S., Haggström, O., & Nerman, O. (1992). *EMPHT — A program for fitting phase-type distributions* (Studies in Statistical Quality Control and Reliability, Mathematical Statistics). Sweden: Chalmers University and University of Göteborg.

Johnson, M. A. (1993a). Selecting parameters of phase distributions: Combining nonlinear programming, heuristics, and Erlang distributions. *ORSA Journal on Computing, 5*, 69–83.

Johnson, M. A. (1993b). An empirical study of queueing approximations based on phase-type distributions. *Stochastic Models, 9*, 531–561.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: An algorithmic approach*. Baltimore: The Johns Hopkins University Press (Reprinted by Dover Publications, 1994).

O'Cinneide, C. A. (1990). Characterization of phase-type distributions. *Stochastic Models, 6*, 1–57.

Pagano, M. E., & Neuts, M. F. (1981). Generating random variates from a distribution of phase type. In T. I. Oren, C. M. Delfosse, & C. M. Shub (Eds.), *1981 Winter simulation conference proceedings* (pp. 381–387). New Jersey: Institute of Electrical and Electronics Engineers.

Schmickler, L. (1992). MEDA: Mixed Erlang distributions as phase-type representations of empirical distribution functions. *Stochastic Models, 8*, 131–156.

Piecewise Linear Function

A function that is formed by linear segments or one that approximates a nonlinear function by linear segments.

Pivot Column

The column vector of coefficients associated with the entering basis variable in a simplex method iteration. Also, more generally, the column that contains the pivot element of a Gaussian elimination step or similar process.

See

- ▶ Eta Vector
- ▶ Gaussian Elimination
- ▶ Matrices and Matrix Algebra
- ▶ Pivot Element
- ▶ Pivot Row
- ▶ Simplex Method (Algorithm)

Pivot Element

In the simplex method, the coefficient of the pivot column whose row index corresponds to the basic variable that is to be dropped from the basis. Also, the element of the pivot column in a Gaussian elimination step that is selected to be on the diagonal of the associated upper triangular matrix.

See

- ▶ Eta Vector
- ▶ Gaussian Elimination
- ▶ Matrices and Matrix Algebra
- ▶ Pivot Column
- ▶ Pivot Row
- ▶ Simplex Method (Algorithm)

Pivot Row

The row corresponding to the position of the basic variable that is to be dropped from the basis in a simplex method iteration. In general, the row correspoding to the row position of a pivot element in a Gaussian elimination step.

See

- ▶ Eta Vector
- ▶ Gaussian Elimination
- ▶ Matrices and Matrix Algebra
- ▶ Pivot Column
- ▶ Pivot Element
- ▶ Simplex Method (Algorithm)

Pivot-Selection Rules

In the simplex method, the pivot selection rules determine which variable is to enter the basic solution and which variable is to be dropped. Depending on the solution at hand, the rules are designed to preserve feasibility (nonnegativity) of the solution (primal-simplex method), or to preserve the optimality conditions (dual-simplex method). In either case, the rules attempt to select an entering variable that would cause and improvement in the objective function. These rules are often augmented with anti-degeneracy or anticycling rules, and procedures for maintaining sparsity and numerical accuracy.

See

▶ Bland's Anticycling Rules
▶ Density
▶ Devex Pricing
▶ Linear Programming
▶ Matrices and Matrix Algebra
▶ Perturbation Methods
▶ Simplex Method (Algorithm)

PMF

Probability mass function.

PO

▶ Postoptimal Analysis

Point Stochastic Processes

Igor Ushakov
Qualcomm Inc., San Diego, CA, USA

Introduction

A point process is a stochastic process $\{N(t), t \geq 0\}$, where $N(t) =$ number of occurrences by time t, which describes the appearance of a sequence of instant random events in time. Usually (though not always) intervals between two neighboring events are considered to be independently distributed. A process of this type is called a point process with restricted memory. If times between occurrences are a sequence of independent and identically distributed (i.i.d.) random variables, the point process is called a renewal or recurrent point process. The Poisson process represents a particular case of a renewal process in which the intervals between occurrences are exponentially distributed (Cox and Isham, 1980; Daley and Vere-Jones, 2002, 2007; Franken et al. 1981).

A special type of point process can be formed by two independent subsequences of random variables that alternate, as in the sequence X_1, Y_1, X_2, Y_2,.... Such a process is called an alternating point process, and more specifically, an alternating renewal process if the X and Y subsequences are themselves ordinary renewal processes.

Thinning of a Point Process

In some cases, events are excluded from the point process with a specified probability. For instance, a unit failure leads to a system failure only if several additional random circumstances happen. This exclusion of events is called a thinning procedure. If the thinning procedure results in the (normalized) probability of the event exclusion going to 1, the resulting point process converges to a Poisson process. This statement is reflected in strong terms in Renyi's Limit Theorem and in its generalization made by Yu. K. Belyaev (see Gnedenko et al. 1969). For practical purposes, the result means that if the mean time between neighboring events in the initial recurrent process equals T, and each event is excluded from this process with the probability p close to 1, the resulting process will be a Poisson process with parameter

$$\lambda = \frac{1-p}{T}.$$

The Superposition of Point Processes

The next important statement concerns the superposition of point processes, which is formulated in the Khinchine-Osokov Limit Theorem (Khinchine 1960; Osokov 1956) and later generalized in the Grigelionis-Pogozhev Limit Theorem (Grigelionis 1964; Pogozhev 1964). On a qualitative level, the theorem states that a limiting point process, which is formed by the superposition of independent "infinitesimally rare" point processes, converges to a Poisson process. For instance, if a piece of equipment consists of a large number of blocks and modules, the flow of its failures may well be considered to form a Poisson process. The parameter of this resulting process is expressed as a sum of the parameters of the initial processes, that is, if there are n recurrent processes ($n >> 1$), each of them with mean T_i, then the resulting process will be close to a Poisson process with parameter

$$\lambda = \sum_{1 \le i \le n} \frac{1}{T_i}.$$

As a consequence of these results, the Poisson process plays a role in the theory of stochastic processes that is analogous to that of the normal distribution in general probability and statistical theory.

See

- ▶ Poisson Process
- ▶ Queueing Theory
- ▶ Renewal Process
- ▶ Stochastic Model

References

Cox, D. R., & Isham, V. (1980). *Point processes*. New York: Chapman and Hall.

Daley, D. J., & Vere-Jones, D. (2002). *An introduction to the theory of point processes, volume 1: Elementary theory and methods* (2nd ed.). New York: Springer.

Daley, D. J., & Vere-Jones, D. (2007). *An introduction to the theory of point processes, volume 2: General theory and structure* (2nd ed.). New York: Springer.

Franken, P., König, D., Arndt, U., & Schmidt, V. (1981). *Queues and point processes*. Berlin, Germany: Akademie-Verlag.

Gnedenko, B. V., Belyaev, Y. K., & Solovyev, A. D. (1969). *Mathematical methods of reliability theory*. New York: Academic Press.

Grigelionis, B. I. (1964). Limit theorems for sums of renewal processes. In A. I. Berg, N. G. Bruevich, & B. V. Gnedenko (Eds.), *Cybernetics in the service of communism, vol. 2: reliability theory and queueing theory* (pp. 246–266). Moscow: Energiya.

Khintchine, A. Y. (1960). *Mathematical methods in the theory of queueing*. London: Charles Griffin.

Osokov, G. A. (1956). A limit theorem for flows of similar events. *Theory Probability and Its Applications, 1*, 246–255.

Pogozhev, I. B. (1964). Estimation of deviation of failure flow in multi-use equipment from Poisson Process (Russian). *Cybernetics in Service for Communism* (vol. 2). Moscow: Energiya.

Point-to-Set Map

A function that maps a point of one space into a subset of another.

Poisson Arrivals

Term used when customers coming to a queueing system follow a Poisson process; this also implies that the time between customer arrivals are independent and identical distributed random variables following an exponential distribution with mean equal to the inverse of the Poisson arrival rate.

See

- ▶ Exponential Arrivals
- ▶ Poisson Process
- ▶ Queueing Theory

Poisson Process

A stochastic, renewal-counting point process beginning from time $t = 0$ with $N(0) = 0$ that satisfies the following assumptions is called a Poisson process with rate λ: (1) the probability of one event happening in the interval $(t, t + h]$ is $\lambda h + o(h)$, where $o(h)$ is a function which goes to zero faster than h; (2) the probability of more than one event happening in $(t, t + h]$ is $o(h)$; and (3) events happening in non-overlapping intervals are statistically independent. (Either (1) or (2) can be replaced by: the probability of no event happening in the interval $(t, t + h]$ is $1 - \lambda h + o(h)$). For such a Poisson process, the times between events (renewals) are independent and identically exponentially distributed with mean $1/\lambda$. In Kendall's queueing notation, arrivals following a Poisson process would be represented by "M" as in an $M/G/1$ queue. An important property of Poisson arrival processes in queueing theory is PASTA (Poisson arrivals see time averages).

See

- ▶ Kendall's Notation
- ▶ Markov Chains
- ▶ Markov Processes
- ▶ PASTA
- ▶ Queueing Theory

Politics

Frederic H. Murphy[1], Sidney W. Hess[2] and
Carlos G. Wong-Martinez[3]
[1]Temple University, Philadelphia, PA, USA
[2]Chadds Ford, Philadelphia, PA, USA
[3]Woosong University, Daejeon, Korea

Introduction

Applications of OR/MS to the representation and electoral processes are considered here. The narrower definition of politics is followed, denoting the theory and practice of managing political affairs in a party sense (Webster's New Collegiate Dictionary 1951). In particular, applications to the following are considered:

- Apportionment
- Districting
- Voting methods and logistics, and
- Election analysis

Apportionment

This is the process of equitably assigning a fixed number of legislators to a lesser number of political subdivisions. In the United States, 435 congressional districts must be apportioned to 50 states with each state receiving at least one district. The method of rounding to an integer solution influences the political result.

Balinski and Young (1982) have provided an exceptional mathematical analysis of the issue along with an historical, nontechnical exposition. In 1791, following the first U.S. census, Jefferson and Hamilton proposed alternative methods for apportionment, the method of greatest divisors (take the ratio of every state's population and the largest divisor such that the integer portions of the ratios add up to the number of representatives to allocate,) and the method of greatest remainders (take the population in a political unit, divide by the total population and multiply by the number of seats, allocate the integer portion, allocate the remaining seats in order of the size of the remainders until there are none left). Washington exercised the first presidential veto when he disagreed with Congress' support of Hamilton's method.

Most methods are biased; for example Jefferson's favors the more populated states while the method used in the United States since 1941, the "method of equal proportions" (also known as the Hill or Huntington method) discriminates against them. In this method a multiplier for adding the nth congressperson to a state is constructed by taking the square root of $1/[n(n - 1)]$, $n > 1$. The product of the multipliers and the states' populations are sorted from highest to lowest for all states together. After each state is given one seat, the remaining seats are given to the 385 highest products of the populations and the multipliers. Other methods exhibit the paradox of a state's apportioned number of seats declining as the total number of representatives increases even when all states' populations are unchanged!

Balinski and Young (1982) conclude that there can be no perfect method. However, Senator Daniel Webster promoted a method called "major fractions" (frequently used between 1842 and 1932), which has been felt by many to be preferable. It is simple, and exhibits neither bias nor the population paradox.

Furthermore, Webster's method (find a divisor for the populations of each political unit such that the rounded quotients sum to the total number of legislators to be allocated), is more likely than the other methods to give each state its proportional number of seats, either rounded up or rounded down (Ernst 1994). See Apportionment Politics for detailed descriptions of apportionment methods and examples of the paradoxes that result from the different apportionment methods.

In most countries once the districts are established the candidate with the most votes wins. In Switzerland an alternative approach is taken to ensure that smaller parties are represented, Beroggi (2010). First, seats are allocated to states using major fractions. Second, seats are allocated to parties at the national level using the same method. With these allocations as constraints, the seats in every district are allocated to parties, minimizing the deviations between the real-valued allocation and integer number of seats ultimately given to each party in each district.

Redistricting

This is the process of defining geographic boundaries for the representatives in a political unit such as a city, state, province, or country. Historically, the party controlling the legislature draws districting maps to protect incumbents and increase their party's chances of maintaining control.

In 1962, the Supreme Court required population equality among districts, demanding more careful mapping than the usual prior political process (Baker v. Carr 1962). A variety of techniques to computerize the mapping process appeared. Most approaches incorporated population equality with the additional criteria that each district be:

– Contiguous, a single land parcel,
– Compact, consolidated rather than spread out, and
– Designed without political consideration.

Hess et al. (1965) solved a sequence of transportation linear programs. In each LP, equal population was allocated to trial district centers to minimize total cost. The measure of cost was compactness defined as the second moment of population about its district center. Centroids of the resultant districts became new centers for repeating the linear program. Successive solution of the transportation problems trended to more compactness while maintaining near population equality. Their heuristic handled problems as large as 350 population units by 19 districts. Larger problems were apportioned into smaller ones. This Ford Foundation-supported program was used for districting in at least seven states.

Hojati (1996) used Lagrangian relaxation to determine the center of districts and then the transportation model to assign population units to districts, followed by a capacitated transportation model to rejoin split population units. George et al. (1997) have generalized the transportation LP into a minimum-cost network-flow formulation that permits more flexible objective functions. They demonstrate objective (cost) functions that include penalties for:

– District populations deviating from the average or exceeding some maximum deviation,
– Districts crossing geographic barriers, and
– Changes from prior district boundaries.

The procedure has been applied in preparing New Zealand legislative-district boundaries involving assignment of 35,000 geographic units to 95 Parliamentary districts.

Garfinkel and Nemhauser (1969) developed a tree search algorithm that minimizes compactness while constraining maximum allowable population deviation. Their measure of district compactness is the diameter squared divided by area. Computation speed and capacity limited the problem size to about 50 population units by seven districts.

Nygreen (1988) redistricted Wales by three different solution methods: solving the integer programming formulation directly, using set partitioning (a variant of Garfinkel and Nemhauser's technique), and using implicit enumeration to structure the search of the tree of solutions. Although his example was small, he concluded that the integer programming technique was inferior. He felt problems to about 500 population units by 60 districts could be solved efficiently by set partitioning. Twenty years of computer improvement permit a tenfold larger problem!

All these redistricting techniques require apportioning a problem too large for solution into many smaller and solvable ones. Apportioning first has added benefits: small political subdivisions are more likely to remain intact and district boundaries will more often coincide with political boundaries.

Hess (1971) showed how first apportioning New York legislative seats to groups of counties minimizes the number of counties that must be in more than one district.

Mehrotra et al. (1998) model the problem as a constrained graph-partitioning problem as in Garfinkel and Nemhauser (1969) and develop a specialized branch-and-price based solution methodology rather than use implicit enumeration. Their reason for generating districts and solving the partitioning problem is to guarantee contiguous districts.

They did not work directly with the facility location/p median problem because ensuring contiguity would require an exponential number of constraints as with sub-tour elimination in the traveling salesman problem. Bozkaya et al. (2003) developed a tabu search approach to solving the problem while restricting the search to contiguous districts, again, not representing contiguity directly because of the perceived difficulty of capturing contiguity.

For such a heavily researched problem with so many successful researchers working on it over decades one would not expect an important breakthrough on a problem as difficult as the contiguity problem. However, two approaches represent contiguity directly in a model without any combinatorial explosion. Williams (2002) shows how to enforce contiguity using constraints on trees defined over the primal and dual planar graphs of the districts. Shirabe (2009), building on work by Zoltners and Sinha (1983), imposes contiguity by modeling trees with constraints that require the adjacent nodes connected by a positive flow be in the same district and the root node have an inflow that matches the number of geographic units assigned to the district. Thus, there has been substantial analytic progress in developing usable models for doing districting using integer programming formulations.

Meanwhile, the courts and legislatures have been slow to articulate permissible or required criteria for districting. A multitude of definitions or measures of compactness are available for Court selection, but all suffer from one flaw or another (Young 1988). In the United States “one man, one vote” is still the law of the land. The 1982 Voting Rights Act requires states with histories of racial discrimination to provide a reasonable chance of minority elections (Van Biema 1993). However, the Supreme Court (Shaw v. Hunt 1996) ruled that racial considerations cannot alone justify bizarre shaped districts. While the courts scrutinize the results of districting, they have not yet challenged the process (Browdy 1990), let alone find political gerrymandering to be unconstitutional. Associate Supreme Court Justice Breyer has regretted that the Court failed to take a stand (King 2010).

Political parties have been free to use proprietary software to generate districting plans that would make Governor Gerry blush. Computer services generated over 1,000 plans for Florida alone, making it difficult for the press and public to criticize gerrymandering (Miniter 1992). It is possible to predict when gerrymandering will happen: if only one political party controls the legislature and the politicians control the process without an independent oversight board, the districts will be drawn to the advantage of that party. That is, the process is important in determining the outcome.

The problem with gerrymandered districts after they are drawn is, like pornography, we know it when we see it. However, it is very difficult to define what gerrymandering is in advance. Consequently, any effort to reduce the degree of gerrymandering has to include not only good analytical models but also good governance processes.

Should the courts order an open districting process or bipartisanship necessitate, optimization models and algorithms could provide a viable approach to aid in redrawing representative boundaries (Browdy 1990). Given the unwillingness of politicians to give up the advantages that come from manipulating district boundaries, the likely eventual outcome will be a mix with optimization modeling establishing baselines and politicians making limited adjustments. Designing such a process will be an interesting challenge.

Voting Methods and Logistics

The application of approval voting was pioneered in the election processes of The Institute of Management Sciences (Fishburn and Little 1988). Here, a voter checks off (approves) any number of the candidates on a ballot, from a single one to potentially every one, with the person having the most checks being declared the winner. Regenwetter, and Grofman (1998) confirm the value of approval voting by examining the outcomes of seven elections, one of them being an INFORMS election.

Savas et al. (1972) reduced the number of New York City election districts by locating multiple voting machines at polling places. The City achieved significant cost savings and increased the probability voters would find functioning machines, without a significant increase in voter distance to the polls.

Election Analysis

The literature on OR in elections is sparse. The main roles seem to be in forecasting and game-theoretic analyses of policies. Barkan and Bruno (1972) used allocation techniques and statistical analysis to aid the 1970 California election campaign of Senator Tunney. Their analyses targeted precincts for voter registration and get-out-the-vote efforts. The key to their success was the ability to identify swing precincts by estimating party loyalty. Soberman and Sadoulet (2007) provide a game-theoretic analysis of rules to limit campaign spending.

A great deal of effort has been put into forecasting the outcome of elections. Campbell and Lewis-Beck (2008) survey past work in forecasting U.S. presidential elections and Lewis-Beck (2010) covers European election forecasting. Both of these articles are introductions to special issues on election forecasting, covering the broadly defined approaches of surveys, econometric analyses, and crowd sourcing such as the Iowa Electronic Market where people bet on the outcome and the prices and odds are set as in pari-mutuel betting. See also Kaplan and Barnett (2003).

See

- ▶ Integer and Combinatorial Optimization
- ▶ Linear Programming
- ▶ Location Analysis
- ▶ Transportation Problem

References

Baker v. Carr. (1962). 369 U.S. 186.

Balinski, M. L., & Young, H. P. (1982). *Fair representation. Meeting the ideal of one man, one vote*. New Haven, CT: Yale University Press.

Barkan, J. D., & Bruno, J. E. (1972). Operations research in planning political campaign strategies. *Operations Research, 20*, 925–941.

Beroggi, G. E. G. (2010). When O.R. becomes the law. *OR/MS today, 37*(3), 44–47.

Bozkaya, B., Erkut, E., & Laporte, G. (2003). A tabu search heuristic and adaptive memory procedure for political districting. *European Journal of Operational Research, 144*, 12–26.

Browdy, M. H. (1990). Computer models and post-Bandemer redistricting. *Yale Law Journal, 99*, 1379–1398.

Campbell, J. E., & Lewis-Beck, M. S. (2008). US presidential election forecasting: An introduction. *International Journal of Forecasting, 24*, 189–192.

Ernst, L. R. (1994). Apportionment methods for the house of representatives and the court challenges. *Management Science, 40*, 1207–1227.

Fishburn, P. C., & Little, J. D. C. (1988). An experiment in approval voting. *Management Science, 34*, 555–568.

Garfinkel, R. S., & Nemhauser, G. L. (1969). Optimal political districting by implicit enumeration techniques. *Management Science, 16*, B495–B508.

George, J. A., Lamar, B. W., & Wallace, C. A. (1997). Political district determination using large-scale network optimization. *Socio-Economic Planning Sciences, 31*, 11–28.

Hess, S. W. (1971). One-man one-vote and county political integrity: Apportion to satisfy both. *Jurimetrics Journal, 11*, 123–141.

Hess, S. W., Weaver, J. B., Seigfeldt, H. J., Whelan, J. N., & Zitlau, P. A. (1965). Nonpartisan political redistricting by computer. *Operations Research, 13*, 998–1006.

Hojati, M. (1996). Optimal political districting. *Computers and Operations Research, 23*(12), 1147–1161.

Kaplan, E., & Barnett, A. (2003). A new approach to estimating the probability of winning the presidency. *Operations Research, 51*(1), 32–40.

King, L. (2010). CNN television interview of associate Supreme Court Justice Stephen Breyer, September 15.

Lewis-Beck, M. S. (2010). European election forecasting: An introduction. *International Journal of Forecasting, 26*, 9–10. Intro to special issue on election forecasting in Europe.

Mehrotra, A., Johnson, E. L., & Nemhauser, G. L. (1998). An optimization based heuristic for political districting. *Operations Research, 44*(8), 1100–1114.

Miniter, R. (1992). Running against the computer; Stephen Solarz and the technician-designed congressional district. *The Washington Post*, September 20, C5.

Nygreen, B. (1988). European assembly constituencies for Wales – Comparing of methods for solving a political districting problem. *Mathematical Programming, 42*, 159–169.

Regenwetter, M., & Grofman, B. (1998). Approval voting, Borda winners, and Condorcet winners: Evidence from seven elections. *Management Science, 44*(4), 520–533.

Savas, E. S., Lipton, H., & Burkholz, L. (1972). Implementation of an OR approach for forming efficient districts. *Operations Research, 20*, 46–48.

Shaw v. Hunt. (1996). 116 S. Ct. 1894, 135 L. Ed. 2d 207.

Shirabe, T. (2009). Districting modeling with exact contiguity constraints. *Environment and Planning B: Planning and Design, 36*, 1053–1066.

Soberman, D., & Sadoulet, L. (2007). Campaign spending limits and political advertising. *Management Science, 53*(10), 1521–1532.
Van Biema, D. (1993). Snakes or ladders. *Time, 12*, 30–33.
Webster's New Collegiate Dictionary. (1951). *Politics* (p. 654). New York: Mirriam.
Wikipedia, Apportionment. Retrieved from http://en.wikipedia.org/wiki/Apportionment_%28politics%29
Williams, J. C. (2002). A zero-one programming model for contiguous land acquisition. *Geographical Analysis, 34*(4), 330–349.
Young, H. P. (1988). Measuring the compactness of legislative districts. *Legislative Studies Quarterly, 13*(1), 105–115.
Zoltners, A. A., & Sinha, P. (1983). Sales territory alignment: A review and model. *Management Science, 29*, 1237–1256.

Pollaczek-Khintchine Formula

For the M/G/1 queueing system, with L defined as the steady-state expected number of customers in the system, λ the customer arrival rate, 1/µ the mean service time and σ^2 the variance of the service distribution, the Pollaczek-Khintchine (P-K) (mean-value) formula gives

$$L = \rho + \left(\rho^2 + \lambda^2\sigma^2\right)/[2(1-\rho)]$$

where $\rho = \lambda/\mu$. Sometimes, the formulas for mean queue size, L_q, mean line delay, W_q, and mean system waiting time, W, which can be easily derived from L using Little's formula, are also called the P-K formulas. More generally, there are associated transform relationships giving the generating function of the steady-state number in system (or queue length) and the Laplace transform of the steady-state delay/waiting times in terms of the Laplace transform of the service time distribution, which are referred to as Pollaczek-Khintchine (P-K) transform formulas.

See

- ▶ Queueing Theory

Polling System

Where a single server visits each group of customers (queue) in cyclic order and then polls to see if there is anyone present. If yes, the service facility serves those customers under such rules as gated (serve only those present when polled) or exhaustive (serve until no customers are left at the location).

See

- ▶ Networks of Queues
- ▶ Queueing Theory

Polyhedron

The solution space defined by the intersection of a finite number of linear constraints, an example of which is the solution space of a linear-programming problem. Such a space is convex.

See

- ▶ Convex Set
- ▶ Linear Programming

Polynomial Hierarchy

A general term used to refer to all of the various computational complexity classes.

See

- ▶ Computational Complexity

Polynomially Bounded (–Time) Algorithm (Polynomial Algorithm)

An algorithm for which it can be shown that the number of steps required to find a solution to a problem is bounded by a polynomial function of the problem's data.

See

- ▶ Computational Complexity
- ▶ Exponential-Bounded (–Time) Algorithm

Polynomial-Time

- ▶ Computational Complexity

Polynomial-Time Reductions and Transformations

- ▶ Computational Complexity

POMDP

- ▶ Partially Observed Markov Decision Processes

Population-based Search Methods

Optimization search methods that propagate a population of solutions from iteration to iteration of the algorithm, generally using evolutionary operators. Examples include genetic algorithms, ant colony optimization, and particle swarm optimization.

See

- ▶ Evolutionary Algorithms
- ▶ Genetic Algorithms
- ▶ Particle Swarm Optimization
- ▶ Swarm Intelligence

Portfolio Analysis

- ▶ Financial Engineering
- ▶ Portfolio Theory: Mean-Variance Model

Portfolio Theory: Mean-Variance Model

John L. G. Board[1], Charles M. S. Sutcliffe[2] and
William T. Ziemba[3,4]
[1]Henley Business School, University of Reading,
Reading, UK
[2]University of Reading, Reading, UK
[3]University of British Columbia, Vancouver,
British Columbia, Canada
[4]Oxford University, Oxford, UK

Introduction

The heart of the portfolio problem is the selection of an optimal set of investment assets by rational economic agents. Although elements of portfolio problems were discussed in the 1930s and 1950s by Allais, De Finetti, Hicks, Marschak and others, the first formal specification of such a selection model was by Markowitz (1952, 1959), who defined a mean-variance model for calculating optimal portfolios. Following Tobin (1958, 1965), Sharpe (1970) and Roll (1972), this portfolio selection model may be stated as

$$\begin{aligned} \text{Minimize} \quad & \boldsymbol{x}'\boldsymbol{V}\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{x}'\boldsymbol{r} = r_p \\ & \boldsymbol{x}'\boldsymbol{e} = 1 \end{aligned} \tag{1}$$

where $\boldsymbol{x}$ is a column vector of investment proportions in each of the risky assets, $\boldsymbol{V}$ is a positive semi-definite variance-covariance matrix of asset returns, $\boldsymbol{r}$ is a column vector of expected asset returns, r_p is the investor's target rate of return and $\boldsymbol{e}$ is a column unit vector. An explicit solution for the problem can be found using the procedures described in Merton (1972), Ziemba and Vickson (1975), or Roll (1972).

Restrictions on short selling can be modeled by augmenting (1) by the constraints

$$\boldsymbol{x} \geq \boldsymbol{0} \tag{2}$$

where $\boldsymbol{0}$ is a column vector of zeros. The problem now becomes a classic example of quadratic mathematical programming; indeed, the development of the portfolio problem coincided with early developments in nonlinear programming. Formal investigations of the properties of both formulations, and variants, appear in Szegö (1980), Huang and Litzenberger (1988), and the references above.

The Use of Mean and Variance

The economic justification for this model is based on the von Neumann-Morgenstern expected utility results, discussed in this context by Markowitz (1959). The model can also be viewed in terms of consumer choice theory together with the characteristics model developed by Lancaster (1971). His argument is that goods purchased by consumers seldom yield a single, well-defined service; instead, each good may be viewed as a collection of attributes, each of which gives the consumer some benefit (or disbenefit). Thus, preference is defined over those characteristics embodied in a good rather than over the good itself. The analysis focuses attention on the attributes of assets rather than on the assets per se. This requires the assumption that utility depends only on the characteristics. With k characteristics, C_k,

$$U = f(W) = g(C_l, \ldots, C_k)$$

where U and W represent utility and wealth. Modeling too few characteristics will yield apparently false empirical results. Clearly, the benefits of this approach increase as the number of assets rises relative to the number of characteristics. The objects of choice are the characteristics $C_1, \ldots, C_k$. In portfolio theory, these are taken to be payoff (return) and risk.

At Markowitz's suggestion, when dealing with choice among risky assets, payoff is measured as the expected return of the distribution of returns, and risk by the standard deviation of returns. Apart from minor exceptions (Ziemba and Vickson 1975), this pair of characteristics form a complete description of assets which is consistent with expected utility theory in only two cases: assets have normal distributions, or investors have quadratic utility of wealth functions. The adequacy of these assumptions has been investigated by a number of authors (e.g., Borch 1969; Feldstein 1969; Tsiang 1972). Although returns have been found to be non-normal and the quadratic utility has a number of objectionable features (not least diminishing marginal utility of wealth for high wealth), several authors demonstrate approximation results that are sufficient for mean-variance analysis (Samuelson 1970; Ohlson 1975; Levy and Markowitz 1979).

A number of authors, including Markowitz (1959), consider alternatives to the variance and suggest the use of the semi-variance. This suggestion has been extended into workable portfolio selection rules. Fama (1971) and Tsiang (1973) have argued the usefulness of the semi-interquartile range as a measure of risk. Kraus and Litzenberger (1976) and others have examined the effect of preferences defined in terms of the third moment, which allows investor choice in terms of skewness. Kallberg and Ziemba (1979, 1983) show that risk aversion preferences are sufficient to determine optimal portfolio choice if assets have normally distributed returns whatever the form of the assumed, concave, utility function.

Solution of Portfolio Selection Model

In the absence of short sales restrictions, (1) can be rewritten as

$$\begin{aligned}\text{Minimize } \boldsymbol{L} = \tfrac{1}{2}\boldsymbol{x}'\boldsymbol{V}\boldsymbol{x} - \lambda_1\left(\boldsymbol{x}'\boldsymbol{r} - r_p\right) \\ - \lambda_2(\boldsymbol{x}'\boldsymbol{e} - 1)\end{aligned} \tag{3}$$

The first-order conditions are

$$\boldsymbol{V}\boldsymbol{x} = \lambda_1\boldsymbol{r} + \lambda_2\boldsymbol{e}$$

which shows that, for any efficient $\boldsymbol{x}$, there is a linear relation between expected returns $\boldsymbol{r}$ and their covariances, $\boldsymbol{V}\boldsymbol{x}$.

Solving for $\boldsymbol{x}$:

$$\boldsymbol{x} = \lambda_1\boldsymbol{V}^{-1}\boldsymbol{r} + \lambda_2\boldsymbol{V}^{-1}\boldsymbol{e} = \boldsymbol{V}^{-1}[\boldsymbol{r}\ \boldsymbol{e}]\boldsymbol{A}^{-1}\left[r_p 1\right]' \tag{4}$$

where

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}'\boldsymbol{V}^{-1}\boldsymbol{r} & \boldsymbol{r}'\boldsymbol{V}^{-1}\boldsymbol{e} \\ \boldsymbol{r}'\boldsymbol{V}^{-1}\boldsymbol{e} & \boldsymbol{e}'\boldsymbol{V}^{-1}\boldsymbol{e} \end{bmatrix}$$

Substituting (4) into the definition of portfolio variance, $\boldsymbol{x}'\boldsymbol{V}\boldsymbol{x}$, yields

$$\begin{aligned}\boldsymbol{V}_p &= \left[r_p 1\right]\boldsymbol{A}^{-1}\left[r_p 1\right]', \text{ and} \\ \boldsymbol{S}_p &= \left[\frac{cr_p^2 - 2br_p + a}{ac - b^2}\right]^{1/2}\end{aligned} \tag{5}$$

where $\boldsymbol{V}_p$ and $\boldsymbol{S}_p$ represent portfolio variance and standard deviation, respectively. This defines the efficient set, which is a hyperbola in mean/standard-deviation space (or a parabola in mean/variance space). The minimum risk is at $S_{\min} = c^{1/2}$ and $r_{\min} = b/c$ (both strictly positive). Rational risk averse investors will hold portfolios lying on this boundary with $r \geq r_{\min}$.

Each efficient portfolio, p, has an orthogonal portfolio z (i.e., such that $\mathrm{Cov}(r_p, r_z) = 0$) with return

$$r_z = \left(a - br_p\right)/\left(b - cr_p\right)$$

Using this, the efficient set degenerates into the straight line tangent to the hyperbola at p which has intercept $\boldsymbol{r}_z$,

$$\boldsymbol{r} = \boldsymbol{r}_z + \lambda \boldsymbol{s} \tag{6}$$

where $\boldsymbol{r}$ and $\boldsymbol{s}$ represent vectors of the expected return and risks of efficient portfolios, and $\lambda = \left(r_p - r_z\right)/S_p$ can be interpreted as the additional expected return per unit of risk. This is known as the Sharpe ratio (Sharpe 1966, 1994). Equation (6) shows a two-fund separation theorem, such that linear combinations of only two portfolios are sufficient to describe the entire efficient set.

Under the additional assumptions of homogeneous beliefs (so that all investors perceive the same parameters) and equilibrium, (6) becomes the Capital Market Line. The Security Market Line (i.e., the relationship between expected returns and systematic risk or $\boldsymbol{\beta}$), which is the outcome of the Capital Asset Pricing Model (CAPM), can be derived by pre-multiplying (4) by $\boldsymbol{V}$ and simplifying using the definitions of V_p and r_z:

$$\boldsymbol{r} = r_z \boldsymbol{e} + \left(r_p - r_z\right)\boldsymbol{\beta} \tag{7}$$

where $\boldsymbol{\beta} = \boldsymbol{Vx}/V_p$. If it exists, the risk-free rate of interest may be substituted for r_z (definitionally, the risk-free return will be uncorrelated with the return on all risky assets). Equation (7) then becomes the original CAPM in which expected return is calculated as the risk-free rate plus a risk premium (measured in terms of an asset's covariance with the market portfolio). The CAPM forms one of the cornerstones of modern finance theory and is not appropriately addressed here. Discussion of the CAPM can be found in Huang and Litzenberger (1988) and Ferson (1995), while systematic fundamental and seasonal violations of the theory are presented in Ziemba (1994) and Keim and Ziemba (1999).

Short Selling

The assumption that assets may be sold short (i.e., $x_i < 0$) is justified when the model is used to derive analytical results for the portfolio problem. Also, when considering equilibrium (e.g., the CAPM), none of the short selling constraints should be binding (because in aggregate, short selling must net out to zero). However, significant short selling restrictions do face investors in most real markets. These restrictions may be in the form of absolute prohibition, the extra cost of deposits to back short selling or self imposed controls designed to limit potential losses.

The set of quadratic programming problems to find the efficient frontier when short sales are ruled out can be formulated as either minimizing the portfolio risk for a specified sequence of portfolio returns (r_p) by repeatedly solving (1) and (2), or maximizing the weighted sum of portfolio risk and return for a chosen range of risk-return tradeoff parameters (μ) by repeatedly solving (8) as below. This latter approach has the advantages of locating only points on the efficient frontier and, for evenly spaced increments in μ, locating more points on the efficient frontier where its curvature is greatest:

$$\begin{aligned} &\text{Maximize} \quad \alpha = \boldsymbol{x}'\boldsymbol{Vx} - \mu\left(\boldsymbol{x}'\boldsymbol{r} - r_p\right) \\ &\text{Subject to} \quad \boldsymbol{x} \geq \boldsymbol{0} \\ &\qquad\qquad\quad \boldsymbol{x}'\boldsymbol{e} = 1 \end{aligned} \tag{8}$$

When short sales are permitted, a position (long or short) is taken in every asset, while when short selling is ruled out, the solution involves long positions in only about 10% of the available assets. When short selling is permitted, about half the assets are required to be sold short, often in large amounts, and sometimes in amounts exceeding the initial value of the investment portfolio. Indeed, this is the main activity of 'short seller' funds.

In contrast, most models based on portfolio theory, in particular the CAPM, ignore short selling constraints (Markowitz 1983, 1987). This change is consistent with the development of equilibrium models for which institutional restrictions are inappropriate (and if imposed would not be binding). However, when short selling is permitted, the number of asset return observations is required to exceed the number of assets, while complementary slackness means that this condition need not be met when short selling is ruled out. Computational procedures to solve mean-variance models with various types of constraints, and the optimal combination of safe and risky assets for various utility functions are discussed by Ziemba et al. (1974).

Estimation Problems

The model (1) requires estimates of $\boldsymbol{r}$ and $\boldsymbol{V}$ for the period during which the portfolio is to be held. This estimation problem has been given relatively little attention, and many authors, both practitioners and academics, have used historical values as if they were precise estimates of future values. However, Hodges and Brealey (1973), among others, demonstrate the benefits obtained from even slight improvements on historical data.

Estimation risk can be allowed for either by using different methods to forecast asset returns, variances and covariances, which are then used in place of the historical values in the portfolio model, or by using the historical values in a modified portfolio selection technique (Bawa et al. 1979). Since the portfolio selection model of Markowitz takes these estimates as parametric, there is no theoretical guidance on the estimation method and a variety of methods have been proposed to provide the estimates. The single index market model of Sharpe (1963) has been widely applied in the literature to forecast the covariance matrix. Originally proposed to reduce the computation required by the full model, it assumes a linear relation between stock returns and some measure of the market, $\boldsymbol{r} = \boldsymbol{\alpha} + \boldsymbol{\beta}'\boldsymbol{m} = \boldsymbol{\varepsilon}$ (for market index $\boldsymbol{m}$ and residuals $\boldsymbol{\varepsilon}$). This uses historical estimates of the means and variances. However, the implied covariance matrix is $\boldsymbol{V}_1 = \upsilon_m\boldsymbol{\beta\beta}' + \boldsymbol{V}$, where υ_m is the variance of the index, $\boldsymbol{\beta}$ is a column vector of slope coefficients from regressing each asset on the market index and $\boldsymbol{V}$ is a diagonal matrix of the variances of the residuals from each of these regressions. A number of studies have found that models based on the single index model outperform those based on the full historical method (e.g., Board and Sutcliffe 1994).

The overall mean method, first proposed by Elton and Gruber (1973), is based on the finding that, although historical estimates of means are satisfactory, data are typically not stable enough to allow accurate estimation of the $N(N-1)/2$ covariance terms. The crudest solution is to assume that the correlations between all pairs of assets expected in the next period are equal to the mean of all the historic correlations. An estimate of $\boldsymbol{V}$ can then be derived from this. Elton et al. (1978) compared the overall mean method of forecasting the covariance matrix with forecasts made using historical values, and four alternative versions of the single index model. They concluded that the overall mean model was clearly superior. A simplified procedure for estimating the overall mean correlation appears in Aneja et al. (1989).

Statisticians have shown increasing interest in Bayesian methods (Hodges 1976) and particularly James-Stein estimators (Efron and Morris 1975, 1977; Judge and Bock 1978; Morris 1983). The intuition behind this approach is that returns that are far from the norm have a higher chance of containing measurement error than those close to it. Thus, estimates of returns, based on individual share data, are cross-sectionally 'shrunk' towards a global estimate of expected returns which is based on all the data. Although these estimators have unusual properties, they are generally expected to perform well in large samples.

Jorion (1985, 1986) examined the performance of Bayes-Stein estimation using both simulated and small real data sets and concluded that the Bayes-Stein approach outperformed the use of historical estimates of returns and the covariance matrix. However, Jorion (1991) found that the index model outperformed Stein and historical models. Board and Sutcliffe (1994) applied these and other methods to large data sets. They found that, in contrast to earlier studies, the relative performance of Bayes-Stein was mixed. While it produced reasonable estimates of the mean returns vector, there were superior methods (e.g., use of the overall mean) for estimating the covariance matrix when short sales were permitted. They also found that, when short sales were prohibited, actual portfolio performance was clearly improved, although there was little to choose between the various estimation methods.

An alternative approach is to try to control for errors in the parameter estimates by imposing additional constraints on (1). Clearly, ex-ante the solution to such a model cannot dominate (1), however, ex-post, dominance might emerge (i.e., what seems, in advance, to be an inferior portfolio might actually perform better than others). The argument is that adding constraints to (1) to impose lower bounds (i.e., prohibiting short sales) and/or upper bounds (forcing diversification) can be used as an ad hoc method of avoiding the worst effects of estimation risk. Of course, extreme, but possibly desirable, corner solutions will also be excluded by this technique. Cohen and Pogue (1967) imposed upper bounds of 2.5% on any asset. Board and Sutcliffe (1988) studied the effects of placing upper bounds on the investment proportions, which may be interpreted as a response to estimation risk. Using historical forecasts of returns and the covariance matrix, and with short sales excluded, they found that forcing diversification leads to improved actual performance over the unconstrained model. Hensel and Turner (1998) have also studied adjusting the inputs and outputs to improve portfolio performance.

Chopra and Ziemba (1993), following the work of Kallberg and Ziemba (1984), showed that errors in the mean values have a much greater effect than errors in the variances, which are in turn more important than errors in the covariances. Their simulations show errors of the order of 20 to 2 to 1. This quantifies the earlier findings and stresses the importance of having good estimates of the asset means.

Another approach is to use fundamental analysis to provide external information to modify the estimates (Hodges and Brealey 1973). Clearly, among the simplest external data to add are the seasonal (e.g., turn of the year, and month and weekend) effects that have been found in most stock markets around the world. Incorporation of these into the parameter estimates can substantially improve the performance of the model. Ziemba (1994) demonstrated the benefits of factor models to estimate the mean returns.

Concluding Remarks

Only the single period mean-variance portfolio theory model has been considered here. Most of the extensions to multi-period models assume frictionless capital markets, which require the solution of a sequence of instantaneous mean-variance models in which the existence of transactions costs adds enormously to the complexity of the problem. Surveys covering dynamic portfolio theory appear in Constantinides and Malliaris (1995), Ziemba and Vickson (1975), Huang and Litzenberger (1988), and Ingersoll (1987); see also Ziemba and Mulvey (1998).

See

- ▶ Banking
- ▶ Financial Engineering
- ▶ Financial Markets
- ▶ Linear Programming
- ▶ Nonlinear Programming
- ▶ Quadratic Programming

References

Aneja, Y. P., Chandra, R., & Gunay, E. (1989). A portfolio approach to estimating the average correlation coefficient for the constant correlation model. *Journal of Finance, 44*, 1435–1438.

Bawa, V. S., Brown, S. J., & Klein, R. W. (1979). *Estimation risk and optimal portfolio choice*. Amsterdam: North Holland.

Board, J. L. G., & Sutcliffe, C. M. S. (1988). Forced diversification. *Quarterly Review Economics and Business, 28*(3), 43–52.

Board, J. L. G., & Sutcliffe, C. M. S. (1994). Estimation methods in portfolio selection and the effectiveness of short sales restrictions: UK evidence. *Management Science, 40*, 516–534.

Borch, K. (1969). A note on uncertainty and indifference curves. *Review Economic Studies, 36*, 1–4.

Chopra, V. R. (1993). Improving optimization. *Journal of Investing, 2*, 51–59.

Chopra, V. R., & Ziemba, W. T. (1993). The effect of errors in means, variances and covariances on optimal portfolio choice. *Journal of Portfolio Management, 19*(2), 6–13.

Cohen, K. J., & Pogue, J. A. (1967). An empirical evaluation of alternative portfolio selection models. *Journal of Business, 40*, 166–193.

Constantinides, G., & Malliaris, G. (1995). Portfolio theory. In R. A. Jarrow, V. Maksimovic, & W. T. Ziemba (Eds.), *Handbook of finance*. Amsterdam: North-Holland.

Efron, B., & Morris, C. (1975). Data analysis using stein's estimator and its generalizations. *Journal of American Statistical Association, 70*, 311–319.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236*(5), 119–127.

Elton, E. J., & Gruber, M. J. (1973). Estimating the dependence structure of share prices: Implications for portfolio selection. *Journal of Finance, 28*, 1203–1232.

Elton, E. J., Gruber, M. J., & Urich, T. J. (1978). Are betas best? *Journal of Finance, 33*, 1375–1384.

Fama, E. (1971). Risk, return and equilibrium. *Journal of Political Economy, 79*, 30–55.

Fama, E. F. (1976). *Foundations of finance*. Oxford: Basil Blackwell.

Feldstein, M. (1969). Mean variance analysis in the theory of liquidity preference and portfolio selection. *Review Economic Studies, 36*, 5–12.

Ferson, W. (1995). Theory and testing of asset pricing models. In Jarrow, Maksimovic, & Zeimba (Eds.), *Handbook of finance*. Amsterdam: North-Holland.

Hensel, C. R., & Turner, A. L. (1998). Making superior asset allocation decisions: A practitioner's guide. In Ziemba & Mulvey (Eds.), *Worldwide asset and liability modelling* (pp. 62–83). Cambridge: Cambridge University Press.

Hodges, S. D. (1976). Problems in the application of portfolio selection. *Omega, 4*, 699–709.

Hodges, S. D., & Brealey, R. A. (1973). Portfolio selection in a dynamic and uncertain world. *Journal of Financial Analysts, 29*, 50–65.

Huang, C. F., & Litzenberger, R. H. (1988). *Foundations for financial economics*. Amsterdam: North-Holland.

Ingersoll, J. (1987). *Theory of financial decision making*. Lanham: Rowman & Littlefield.

Jarrow, R., Maksimovic, V., & Ziemba, W. T. (Eds.). (1995). *Finance*. Amsterdam: North-Holland.

Jobson, J. D., & Korkie, B. (1981). Putting markowitz theory to work. *Journal of Portfolio Management, 7*, 70–74.

Jobson, J. D., Korkie, B., & Ratti, V. (1979). Improved estimation for Markowitz portfolios using James-Stein type estimators. *Proceedings Business Economics and Statistics Section, American Statistical Association, 41*, 279–284.

Jorion, P. (1985). International portfolio diversification with estimation error. *Journal of Business, 58*, 259–278.

Jorion, P. (1986). Bayes-Stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis, 21*, 279–292.

Jorion, P. (1991). Bayesian and CAPM estimators of the means: Implications for portfolio selection. *Journal of Banking and Finance, 15*, 717–727.

Judge, G. G., & Bock, M. E. (1978). *The statistical implications of pre-test and stein-rule estimators in econometrics*. Amsterdam: North-Holland.

Kallberg, J. G., & Ziemba, W. T. (1979). On the robustness of the Arrow-Pratt risk aversion measure. *Economics Letters, 2*, 21–26.

Kallberg, J. G., & Ziemba, W. T. (1983). Comparison of alternative utility functions in portfolio selection. *Management Science, 29*, 1257–1276.

Kallberg, J. G. & Ziemba, W. T. (1984). Mis-specification in Portfolio Selection Problems. In G. Bamberg & K. Spremann, (Eds.)., Risk and Capital : Lecture Notes In Economic and Mathematical Systems, Springer-Verlag, New York.

Keim, D. B., & Ziemba, W. T. (Eds.). (1999). *Security market imperfections in worldwide equity markets*. Cambridge: Cambridge University Press.

Kraus, A., & Litzenberger, R. F. (1976). Skewness preference and the valuation of risk assets. *Journal of Finance, 31*, 1085–1100.

Lancaster, K. (1971). *Consumer demand: a new approach*. New York: Columbia University Press.

Levy, H. (1969). A utility function depending on the first three moments. *Journal of Finance, 24*, 715–719.

Levy, H., & Markowitz, H. (1979). Approximating executed utility by a function of mean and variance. *American Economic Review, 69*, 308–317.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance, 7*, 77–91.

Markowitz, H. M. (1959). *Portfolio selection: Efficient diversification of investments*. New Haven: Yale University Press.

Markowitz, H. M. (1983). Nonnegative or not non-negative: A question about CAPMs. *Journal of Finance, 38*, 283–295.

Markowitz, H. M. (1987). *Mean-variance in portfolio choice and capital Markets*. Oxford: Blackwell.

Merton, R. C. (1972). An analytic derivation of the efficient portfolio frontier. *Journal of Financial and Quantitative Analysis, 7*, 1851–1872.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of American Statistical Association, 78*, 47–55.

Ohlson, J. (1975). Asymptotic validity of quadratic utility as the trading interval approaches zero. In W. T. Ziemba & R. G. Vickson (Eds.), *Stochastic optimization models in finance*. New York: Academic.

Roll, R. (1972). A critique of the asset pricing theory's tests. *Journal of Financial Economics, 4*, 129–176.

Samuelson, P. (1970). The fundamental approximation theorem of portfolio analysis in terms of means variances and higher moments. *Review of Economic Studies, 37*, 537–542.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management science; Operations research (OR), Practice of Management Science, 9*, 277–293.

Sharpe, W. F. (1966). Mutual fund performance. *Journal of Business, 39*, 119–138.

Sharpe, W. F. (1970). *Portfolio theory and capital markets*. New York: McGraw-Hill.

Sharpe, W. F. (1994). The sharpe ratio. *Journal of Portfolio Management, Fall*, 59–68.

Szegö, G. P. (1980). *Portfolio theory, with application to bank assess management*. New York: Academic.

Tobin, J. (1958). Liquidity preference as behaviour towards risk. *Review of Economic Studies, 26*, 65–86.

Tobin, J. (1965). The theory of portfolio selection. In F. H. Hahn & F. P. R. Brechling (Eds.), *The theory of interest rates* (International Economic Association, pp. 3–51). London: Macmillan.

Tsiang, S. (1972). The rationale of the mean standard deviation analysis, skewness preference and the demand for money. *American Economic Review, 62*, 354–371.

Tsiang, S. (1973). Risk, return and portfolio analysis: Comment. *Journal of Political Economy, 81*, 748–751.

Ziemba, W. T. (1994). World wide security market regularities. *European Journal of Operational Research, 74*, 198–229.

Ziemba, W. T., & Mulvey, J. M. (Eds.). (1998). *World-wide asset and liability modelling*. Cambridge: Cambridge University Press.

Ziemba, W. T., Parkan, C., & Brooks-Hill, F. J. (1974). Calculation of investment portfolios with risk free borrowing and lending. *Management Science, 21*, 209–222.

Ziemba, W. T., & Vickson, R. G. (Eds.). (1975). *Stochastic optimization models in finance*. New York: Academic.

POS

Point of sale.

See

▶ Retailing

Postoptimal Analysis

The study of how a solution changes with respect to (usually) small changes in the problem's data. In particular, this term is applied to the sensitivity analysis and parametric analysis of a solution to a linear-programming problem.

See

▶ Linear Programming
▶ Parametric Programming
▶ Sensitivity Analysis

Posynomial Programming

▶ Geometric Programming

Power Model

▶ Learning Curves

PP

▶ Parametric Programming

PPB(S)

Planning-programming-budgeting (system).

See

▶ Cost Analysis
▶ Military Operations Research

Practice of Operations Research and Management Science

Hugh J. Miser
Farmington, CT, USA

Introduction

The practice of OR/MS here will mean using the appropriate models, tools, techniques, and craft skills of these sciences to understand the problems of people/machine/nature systems with a view toward ameliorating these problems, possibly by new understandings, new decisions, new procedures, new structures, or new policies. Such practice calls for a suitable form of professionalism in dealing not only with the phenomena of the problem situation but also with the persons with relevant responsibilities, as well as other parties at interest.

OR/MS as a Science

Following Ravetz (1971), science in general may be described as "craft work operating on intellectually constructed objects," each object defining a class. Scientific work is thus aimed at establishing new properties of these objects and verifying that they reflect the reality of the classes of phenomena that they represent (Miser 1993). This description has four implications:

1. The intellectual objects – that OR/MS workers usually call models – are created by the imagination, informed by earlier knowledge of the phenomena and objects that have described them successfully, as well as innovative ideas or new evidence from reality.
2. There is a continuing reference to the phenomena of reality.
3. Scientific inquiry then becomes the search for new properties of the classes both by manipulating the objects and seeking new evidence from reality as a basis for revising them.
4. The new properties deduced from the objects – or models – must then be compared with the appropriate aspects of the phenomena of reality.

It is essential to observe that the different sciences – such as physics, biology, or OR/MS – are distinguished, not by their methods, techniques, or models (many of which are widely shared among the sciences), but by the portions of reality in which they are undertaking to understand, explain, and solve problems (Kemeny 1959).

Within the framework established by this conception, it is convenient to distinguish three classes of problems, depending on their goals: to paraphrase Ravetz (1971), scientific problems (where the goal of the work is to establish new properties of the objects of inquiry, and the ultimate function is to achieve knowledge in its field); technical problems (those where the function to be performed specifies the problem); and practical problems (where the goal of the task is to serve or achieve some human purpose and the problem is brought into being by recognizing a problem situation in which some aspect of human welfare should be improved).

Against this background, practice can be recognized as the activity centered on practical problems, even while noting that to solve a practical problem often involves solving technical problems, and, when the basic phenomena underlying a problem situation are not understood, solving scientific problems in order to have the models needed for understanding the practical problem. It is also important to note that this view of science includes work on all three classes of problems within the conception of science as a whole. (For a more extended summary of Ravetz's view of science, see Miser and Quade 1988).

The Context of OR/MS

Since sciences are distinguished by their fields of inquiry, it is important to describe this context for OR/MS if it is to be differentiated from other sciences. In this endeavor the OR/MS community has not reached any sort of brief consensus, so what is said here must be regarded as a personal view, based in part on the literature and in part on personal experience.

While OR/MS deals with systems involving people, elements of nature, and machines (where this last term is intended to include not only artifacts but also laws, standard procedures, common behaviors, and social structures and customs), attempts to take the concept of system beyond this primitive statement as the basis for describing the context of OR/MS have, however, not proved fruitful.

The concept of an action program (Boothroyd 1978) is more useful: a function, operation, or response that is related to and given coherence by a human objective, need, or problem, together with the system of people, equipment, portion of nature, organizational elements, and management or social structure involved.

It is easy to see that an element in an action program may also have membership in other action programs; for example, an executive in one may also play a role in many others, as may also be the case for a major facility or organization, such as a large corporation or a government. Too, an action program may produce effects on other action programs, both through the cross memberships of elements and by the direct impacts of what it does. (For a more extended summary of Boothroyd's concept, see Miser and Quade 1988).

The practice of OR/MS can then be described as the activity that brings the knowledge and skills of the science of OR/MS to bear on the problems of action programs (Miser 1997). While this brief description will suffice as a basis for the argument here, the reader should be aware of the facts that, while it is quite general and covers most of what OR/MS does in practice now, it not only may not cover all of today's activities of practice but also may become even more incomplete with the passage of time.

The Situations of Practice

While each situation in practice may properly be seen as unique, it is nevertheless possible to describe one that contains elements central to most – if not all – of practice, as follows.

An OR/MS analyst is often consulted when someone with a suitable responsibility in an action program discerns a problem situation that needs improvement. While this responsible person may have diagnosed the problem and even may have a notion about a possible solution, it is commonly the case that the forces actually yielding the source of dissatisfaction lie buried deeply enough to make such a diagnosis questionable, and the preconceived fix inappropriate. Thus, typically it is best for the analyst – or the team of analysts if the problem situation is complex – to approach it with an open mind, and aim to explore it thoroughly before deducing its properties and using them to devise a scheme for ameliorating its undesirable properties.

The analysts may be drawn from two sources:

1. There may be an analysis group inside the organization or action program with which the responsible problem-situation identifier – or client – is associated.
2. Analysts may have to be drawn from outside this organization or action program. In either case, there is abundant experience to support the conclusion that a successful outcome of the practice engagement calls for creating a constructive partnership between the analysis team and the parties at interest in the problem situation, as will be discussed in more detail later.

The Processes of Practice

Figure 1 offers a synoptic view of the elements that may be included in a practice engagement that proceeds from the general unease of a problem situation to the implementation of some policy or course of action and evaluates its effects. Since each situation has its own unique properties, few OR/MS practice engagements follow such a procedure exactly, but it is a common experience for many – if not most – of these elements to occur at some stage of the work.

Formulation – The work begins with a thorough exploration of the problem situation in which the client and his/her action program cooperate. The purpose is to formulate the problem to be addressed, which commonly is quite different from the one originally conceived by the client. Once this is done, and the client has agreed with the analysis team on the problem, it is possible to plan the work to be done. This early work also identifies the values and criteria that should inform the choice of what eventually will be done to ameliorate the client's concerns, sets up the objectives to be sought by the solution, and agrees with the client on the boundaries and constraints that must be observed in devising it.

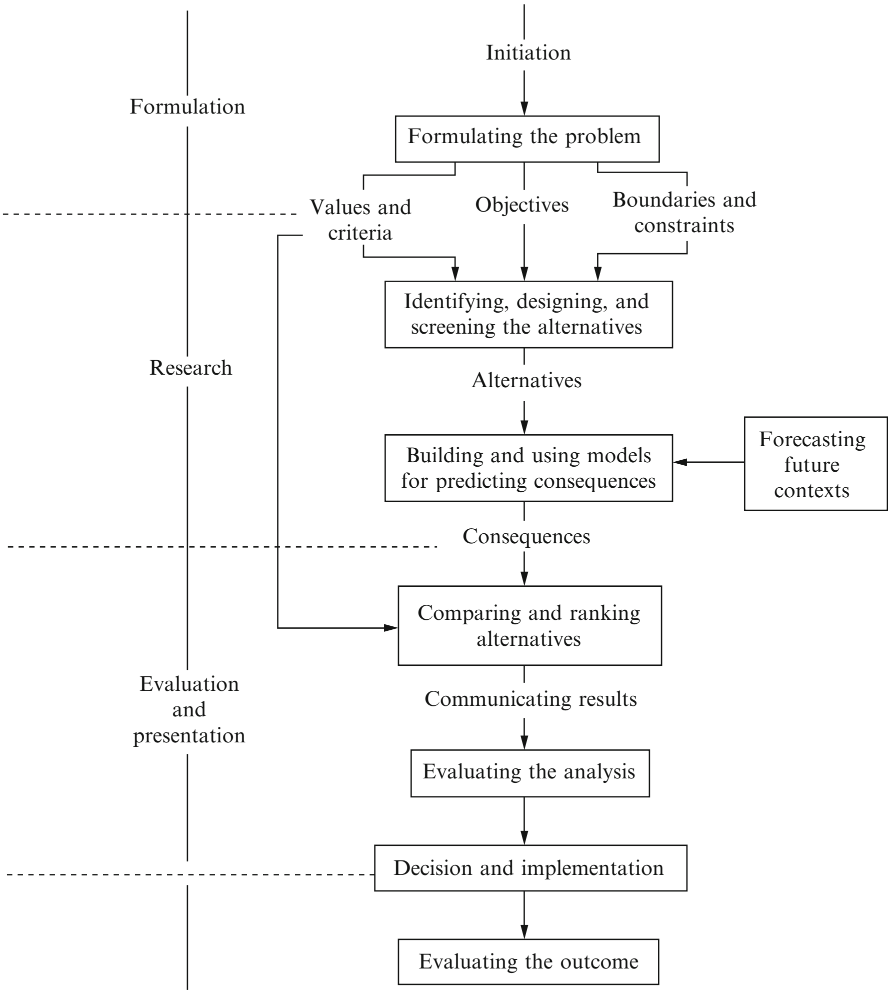
Usually this problem formulation step is one in which the analysts take the lead and work through it in informal cooperation with the client's staff. On occasion, however, it is best for a group consisting of both analysts and members of the client's staff to work together somewhat more formally toward a problem structure. To this end, there are various types of methods (Rosenhead 1996) that can be adapted to these situations. While the results of such a problem-structuring activity are usually a prelude to a more detailed analysis to follow, it sometimes happens that the insights from the group activity shared between the analysts and the client's staff are adequate to show what should be done to ameliorate the problem situation.

Research – This stage extends the information-and data-gathering that began in the formulation stage. The findings that emerge from processing these results allow the analysis team to identify, design, and screen possible alternatives that may help with the problem. Against this background, the analysis team can build models capable of deducing the consequences of adopting each of the alternatives chosen for further investigation within the contexts of possible future conditions.

Evaluation and Presentation – With estimates of the consequences in hand, the analysts may compare – and possibly rank – the alternatives against the criteria chosen earlier in the analysis, plus any new ones that may have emerged during the work. These findings must then be presented to the client and other parties at interest in a way that enables them not only to appreciate the results but also have at least a broad overview of the logic that produced them. These understandings may then enable the client to adopt a suitable policy or course of action.

Although the client, and not the analysts, must decide on what to do and how to carry it out effectively, experience shows that it is very important for the analysis team, or at least analysts who

Practice of Operations Research and Management Science, Fig. 1 Important elements in an OR/MS practice engagement that runs from problem formulation through research and implementation to evaluating the outcome (Source: Miser and Quade (1988), p. 23; reproduced by permission.)


Formulation
Research
Evaluation and presentation
Initiation
Formulating the problem
Values and criteria
Objectives
Boundaries and constraints
Identifying, designing, and screening the alternatives
Alternatives
Building and using models for predicting consequences
Forecasting future contexts
Consequences
Comparing and ranking alternatives
Communicating results
Evaluating the analysis
Decision and implementation
Evaluating the outcome

understand and appreciate what was done, to work cooperatively throughout the implementation stage, as discussed later.

Variations

While it is possible to specify a core diagram of the principal elements of OR/MS practice, it must be admitted immediately that few, if any, such engagements follow this outline exactly. Rather, since each problem situation is different, the analysis activity must be adapted to it. Thus, in studying a series of cases, one sees variations like these:

- Instead of proceeding linearly from the top to the bottom of Fig. 1, the work cycles from intermediate stages back to earlier ones as the progress brings new insights and fresh intermediate results that may prompt reconsideration of the beginning foundations of the work.
- Some work may be aimed more at fleshing out the client's understanding of his situation than prompting him/her to change it significantly, so it may stop at one of the intermediate stages.
- The relative effort expended in the various stages may vary tremendously from case to case: one case may have to expend its major effort in just the information-and data-gathering stage, after which what needs to be done may be fairly apparent without much further analysis. Another case may proceed fairly expeditiously through the outline of Fig. 1 and then have a very long and complicated period of work to achieve what may appear to the outsider to be the implementation of a relatively simple set of proposals.

– In some cases an intermediate stage may dominate the work, owing to such factors as technical difficulty in devising proper models, major uncertainties in forecasting future contexts, complexities of the underlying situation, and so on.

In any case, the procedure specified here as the basis for discussion must be regarded as one that has stitched together the key elements that may enter OR/MS practice to varying extents depending on the peculiarities of the situation being studied.

The Importance of Following Through

The interest of the OR/MS professional, particularly if academically oriented, may flag after the research stage is completed and its results obtained. However, experience shows strongly that to stop there is almost always to waste the earlier effort. Two essential steps must follow: effective communication of the results, and cooperative aid in the implementation process.

Communication – This process, which may not be as appealing to the analyst as the research that preceded it, is nevertheless equally important and deserves great care, since communicating the findings inadequately can vitiate their potential effect, and thus waste the earlier effort. In view of the importance of this step in the OR/MS process, it is surprising that there is no systematic literature describing the skills needed and setting forth how they are best used (for a brief exception see Miser 1985). The discussion will be restricted to these points:

– Few clients will devote a large block of time to such communications, so it is very important to work very hard to condense the principal ideas and findings into as economical a space as possible, whether the form used is oral or written. For example, a top executive may want the key findings presented to him or her in a two-page memorandum or a 20-min briefing. It is perhaps surprising to the uninitiated to see how much important information can be condensed into so small a space, but only if great care is taken to make the best use of it. Graphs and charts accompanying the words can do much to aid this condensation.

– To communicate effectively, the client's vocabulary must be used, with as few technical terms introduced as possible.

– The whole must be focused on the interests of the client or the audience; after a major study many different groups may have to be addressed, and when this is the case the communication instruments must in each case be tailored to the group in view.

– The analysts must be prepared to stand behind their work and to discuss its implications, even those that may go beyond what was done as part of the analysis.

Implementation – It is clear that, if the findings of an OR/MS practice engagement do not find their way into some sort of changed reality, the work is ineffective. Therefore, it is obviously important for the analysis to consider the issue of eventual implementation throughout the work, keeping these points in mind:

1. Since the setting in which the work is being done has properties that will affect how change can be achieved, it is important for the peculiarities of this setting to be kept in mind from the beginning of the analysis. For example, can possible prospective changes be accommodated easily within the existing structure, or will it need to be changed significantly?
2. Since the settings in which OR/MS work is done are so various, it is impossible to stipulate a standard pattern for implementation work. This implies that the findings of the analysis may have to include a prospective implementation structure and program for the decision makers to consider as part of their judgment about the worth of the findings.
3. If the analysis considers different programs of action, the comparisons leading to a preferred choice should consider the relative difficulties of implementation as part of the analysis.
4. The history of analysis records that many well developed and clearly desirable program proposals failed to be implemented because the needed resources either did not exist or could not be made available. Therefore, in conceiving an implementation program as part of the findings of an OR/MS study, it is important to consider its resource requirements, as they will almost surely be an important issue to consider in whether or not to adopt the findings and translate them into action.

No matter how thoroughly the client – or members of his or her staff who participated in the analysis – understand what was found and its prospective implementation, it is a common experience that the implementation process demands the continuing interest and cooperation of the analysis team, or at least some member of it who is able to follow through. The process of change invariably brings up new problems and issues that, wrongly handled, can vitiate the effects of what the original implementation set out to do. Too, these new problems may call for additional complementary analysis that must take account of what was done earlier.

This continuing involvement by analysts in the implementation process may take a variety of forms, ranging from occasional consultation to a continuing direct involvement of a substantial effort over such a long period of time as to make the implementation involvement a more ambitious enterprise than the original analysis (for an example illustrating this last point, see, Mechling 1995).

The roles of the analysts during implementation may include such activities as these:

1. Conducting supplementary analyses when situations arise calling for such work.
2. Helping all concerned keep the goals of the implementation program in sight. (It is all too easy for staff members involved, all of whom have personal in institutional goals in mind, to corrupt what is being done sufficiently that the original goals emerging from the analysis are vitiated.)
3. Proposing changes in the implementation strategy when they are called for by changing circumstance or the appearance of difficulties not foreseen in the beginning.
4. Acting as an on-site agent of persuasion when those directly involved in the implementation program need to have its goals clarified.

In sum, since an effective implementation phase is essential to the success of an OR/MS engagement, analysts should give it as much analytic and administrative importance and support as the analysis phase itself. For further elaboration of these points about implementation, see Tomlinson et al. (1985).

Outcome evaluation – It not infrequently happens that the outcomes of implementations are sufficiently clear to satisfy all concerned. Sometimes, however, in situations complex enough to make the outcomes unclear, it is necessary to conduct additional analysis to estimate the effectiveness of the implemented program or policy. The familiarity of the analysis team with the situation gives it an advantage in conducting such an analysis. However, to eliminate what may appear to be the original analysis team's bias in favor of a good outcome, clients may prefer to call in a new group to conduct such an outcome evaluation.

The Relation Between Analyst and Client

Emerging from a close scrutiny of the relations that should exist between analyst and client for effective cooperation, Schön (1983) advocates a "reflective contract" that works in this way: ". . . in a reflective contract between practitioner and client, the client does not agree to accept the practitioner's authority but to suspend disbelief in it. He agrees to join the practitioner in inquiring into the situation for which the client seeks help; to try to understand what he is experiencing and to make that understanding accessible to the practitioner; to confront the practitioner when he does not understand or agree; to test the practitioner's competence by observing his effectiveness and to make public his questions over what should be counted as effectiveness; to pay for services rendered and to appreciate competence demonstrated. The practitioner agrees to deliver competent performance to the limits of his capacity; to help the client understand the meaning of the professional's advice and the rationale for his actions, while at the same time he tries to learn the meanings his actions have for the client; and to reflect on his own tacit understanding when he needs to do so in order to play his part in fulfilling the contract."

Under this concept for OR/MS work, the client's obligation to share his experience and understanding of the problem situation is often discharged by assigning a member of his staff to work with the analysis team, an arrangement that has many benefits, among which these may be listed: it helps the analysis team identify and gather the information that it needs as a background and basis for its work; it helps the analysts avoid foolish mistakes related to the client's operations; and it acts to keep the client informed of what is emerging from the analysis, which often helps to pre-sell the findings that eventually merge.

Since OR/MS practice may be viewed as a dialogue between analyst and client related to the problem situation and the problem from it that is eventually chosen for analysis, this arrangement serves as a useful continuing conduit for this dialogue, beyond what can be achieved with periodic progress meetings with the client (Miser 1994).

Other practical arrangements between the analysis team and the client to implement Schön's concept of a reflective contract must, of necessity, be evolved in the light of the circumstances peculiar to each engagement. An inhouse analysis group that has been able to achieve a reflective contract with the organization of which it is a part has a special opportunity: it can often identify problem situations that may not yet have been observed by executives in the organization, and thus set to work on them before they grow in size and importance.

How to Learn the Skills of Practice

The OR/MS community has, unfortunately, not evolved a comprehensive epistemology of practice and set it down in easily accessible literature that can be used widely in training courses. Some first steps in this direction for systems analysis, the large-scale efforts that can be thought of as part of OR/MS practice, are taken in Miser and Quade (1985,1988) and Miser (1995); much of what they say can apply equally to OR/MS as a whole. Thus, to learn the needed scientific and craft skills, someone aiming for an OR/MS career must pursue a tripartite program assembled from a variety of sources.

The intellectual basis – The foundation of effective OR/MS practice must be a thorough education in mathematics, with special attention to probability and statistics. Since by now certain models have become associated with OR/MS (as any introductory college textbook makes clear), these should be mastered as well. And a broad view of science with knowledge of other branches is also sure to be helpful.

Beyond a good mathematical and scientific education, however, the potential practitioner must not only be willing but also eager to learn from the problem situation, from the people in it, and from the representatives of other specialties, both practical and intellectual, that may have to be called on to help. As Schön's concept makes clear, to undertake an engagement in practice is to enter a multipartite partnership, and the flow of information must reflect this if the work is to be effective.

Since the action programs that OR/MS practice deals with contain people as essential elements, the analysts must know how to deal effectively and sympathetically with them, since they will enter the problem situation at many levels. In sum, interpersonal skills are an important requisite of good practice.

Familiarity with successful cases – There are by now a great many published accounts of successful cases of OR/MS practice. The journal *Interfaces* specializes in presenting them, and since 1975 has been a treasure-house of such accounts, as well as proven advice about the arts of practice. Assad et al. (1992) accompany a selection of these cases with valuable commentary. For a much wider view, one can consult the "Applications Oriented" section of the *International Abstracts in Operations Research*, the comprehensive abstract journal that has been published since 1961; it will not only exhibit the wide variety of practice being undertaken throughout the world but also identify the many journals and books in which cases appear. Rivett (1994) offers a broad introduction to successful practice based on a lifetime of varied experience.

Apprenticeship – Since the OR/MS community has yet to achieve a widely agreed and centrally documented view of its epistemology of practice, the best way for a person to observe and learn the myriad craft skills of practice is to work with an accomplished and skillful analysis team – in sum, to serve an apprenticeship (Miser and Quade 1985, 1988, offer a substantial body of additional information relating to the craft skills needed for effective OR/MS).

Examples of Good Practice

Since 1975, *Interfaces* has published the finalist papers in the Franz Edelman competition for the best papers on practice each year; there are five or more finalists in each competition. These accounts are an excellent central source of examples of good practice; in recent years tapes of the finalist presentations have also been made available.

There are many other sources of such work – too many to list here; however, both *Operations Research* and the *Journal of the Operational Research Society* contain one or more examples of good practice in each issue, as do the sources mentioned earlier.

See

- ▶ Decision Making and Decision Analysis
- ▶ Ethics in the Practice of Operations Research
- ▶ Field Analysis
- ▶ Implementation of OR/MS in the Public Sector
- ▶ Problem Structuring Methods
- ▶ Systems Analysis

References

Assad, A. A., Wasil, E. A., & Lilien, G. L. (1992). *Excellence in management science practice: A readings book*. New Jersey: Prentice Hall.

Boothroyd, H. (1978). *Articulate intervention*. London: Taylor and Francis.

Kemeny, J. G. (1959). *A philosopher looks at science*. New York: Van Nostrand Reinhold.

Mechling, J. E. (1995). Implementing innovative work schedules in the New York City sanitation department. In H. J. Miser (Ed.), *Handbook of systems analysis: Cases* (pp. 153–196). Chichester, UK: Wiley.

Miser, H. J. (1985). The practice of systems analysis. In H. J. Miser & E. S. Quade (Eds.), *Handbook of systems analysis: Overview of uses, procedures, applications, and practice* (pp. 287–326). Chichester, UK: Wiley.

Miser, H. J. (1993). A foundational concept of science appropriate for validation in operational research. *European Journal of Operational Research*, *66*, 204–215.

Miser, H. J. (1994). Systems analysis as dialogue: An overview. *Technological Forecasting and Social Change*, *45*, 299–306.

Miser, H. J. (Ed.). (1995). *Handbook of systems analysis: Cases*. Chichester, UK: Wiley.

Miser, H. J. (1997). The easy chair: Is it possible to have a good definitional description of operations research and management science? *Interfaces*, *27*(6), 16–21.

Miser, H. J., & Quade, E. S. (Eds.). (1985). *Handbook of systems analysis: Overview of uses, procedures, applications, and practice*. Chichester, UK: Wiley.

Miser, H. J., & Quade, E. S. (Eds.). (1988). *Handbook of systems analysis: Craft issues and procedural choices*. Chichester, UK: Wiley.

Ravetz, J. R. (1971). *Scientific knowledge and its social problems*. Oxford: Oxford University Press (Reprinted 1996, New Brunswick, New Jersey: Transaction Publishers.).

Rivett, P. (1994). *The craft of decision modelling*. Chichester, UK: Wiley.

Rosenhead, J. (1996). What's the problem: An introduction to problem structuring methods. *Interfaces*, *26*(6), 117–131.

Schön, D. H. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Tomlinson, R., Quade, E. S., & Miser, H. J. (1985). Implementation. In H. J. Miser & E. S. Quade (Eds.), *Handbook of systems analysis: Overview of uses, procedures, applications, and practice* (pp. 249–280). Chichester, UK: Wiley.

Precedence Diagramming

A graphic analysis of a project plan in which the nodes are the work activities (or tasks) and are connected by arrows. Relationships among tasks are designated as start-to-start, start-to-finish, and finish-to-finish, which eliminates the use of dummy arrows.

See

- ▶ Network Planning

Predictive Model

A model used to predict the future course of events and as an aid to decision making.

See

- ▶ Decision Problem
- ▶ Descriptive Model
- ▶ Mathematical Model
- ▶ Model
- ▶ Normative Model
- ▶ Prescriptive Model

Preemption

Concept having to do with how priorities are treated. In queueing theory, this means that an arriving higher priority customer pushes a lower one out of service because the newcomer has higher priority; service of the preempted customer later can either continue from the point of its interruption (preemptive resume queue discipline) or start totally anew. In goal programming problem, it is a statement that stipulates the ordering of the goals, so that a solution that satisfies the priority k goal is always to be preferred to solutions that satisfy the lower priority goals $k + 1$,....

See

- ▶ Goal Programming
- ▶ Queueing Theory

Preemptive Priorities

- ▶ Goal Programming
- ▶ Preemption
- ▶ Queueing Theory

Preference Theory

James S. Dyer[1] and Jianmin Jia[2]
[1]The University of Texas at Austin, Austin, TX, USA
[2]The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China

Introduction

Preference theory studies the fundamental aspects of individual choice behavior, such as how to identify and quantify an individual's preferences over a set of alternatives and how to construct appropriate preference representation functions for decision making. An important feature of preference theory is that it is based on rigorous axioms which characterize individual's choice behavior. These preference axioms are essential for establishing preference representation functions, and provide the rationale for the quantitative analysis of preference. Preference theory provides the foundation for economics and the decision sciences. A basic topic of microeconomics is the study of consumer preferences and choices (Kreps 1990). In decision analysis and operations research, knowledge about the decision maker's preference is necessary to establish objective (or preference) functions that are used for evaluating alternatives. Different decision makers usually have different preference structures, which may imply different objective functions for them. Preference studies can also provide insights into complex decision situations and guidance for simplifying decision problems. The basic categories of preference studies can be divided into characterizations of preferences under conditions of certainty or risk and over alternatives described by a single attribute or by multiple attributes. This article begins with the introduction of basic preference relations and then discusses preference representation under certainty and under risk. A preference representation function under certainty will be referred to as a value function, where as a preference representation function under risk will be referred to as a utility function.

Basic Preference Relations

Preference theory is primarily concerned with properties of a binary preference relation $>_p$ on a choice set $\boldsymbol{X}$, where $\boldsymbol{X}$ could be a set of commodity bundles, decision alternatives, or monetary gambles. For example, an individual might be presented with a pair of alternatives, say x and y (e.g., two cars), and asked how they compare (e.g., do you prefer x or y?). If the individual says that x is preferred to y, then write $x >_p y$, where $>_p$ means strict preference. If the individual states that he or she is indifferent between x and y, then this preference is represented as $x \sim_p y$. Alternatively, define $\sim_p$ as the absence of strict preference, i.e., not $x >_p y$ and not $y >_p x$. If it is not the case that $y >_p x$, then write $x \geq_p y$, where $\geq_p$ represents a weak preference (or preference-indifference) relation. Also define $\geq_p$ as the union of strict preference $>_p$ and indifference $\sim_p$ i.e., both $x >_p y$ and $x \sim_p y$.

Preference studies begin with some basic assumptions (or axioms) of individual choice behavior. First, it seems reasonable to assume that an individual can state preference over a pair of alternatives without contradiction, i.e., the individual cannot strictly prefer x to y and y to x simultaneously. This leads to the following definition for preference asymmetry: preference is asymmetric if there is no pair x and y in $\boldsymbol{X}$ such that $x >_p y$ and $y >_p x$.

Asymmetry can be viewed as a criterion of preference consistency. Furthermore, if' an individual makes the judgment that x is preferred to y, then he or she should be able to place any other alternative z somewhere on the ordinal scale determined by the following: either better than y, or worse than x, or both. Formally, define negative transitivity by saying that preferences are negatively transitive if given $x >_p y$ in $\boldsymbol{X}$ and any third element z in $\boldsymbol{X}$, it follows that either $x >_p z$ or $z >_p y$, or both.

If the preference relation $>_p$ is asymmetric and negatively transitive, then it is called a weak order. The weak order assumption implies some desirable properties of a preference ordering, and is a basic assumption in many preference studies. If the preference relation $>_p$ is a weak order, then the associated indifference and weak preference relationships are well behaved. The following results summarize some of these.

If strict preference $>_p$ is a weak order, then

1. strict preference $>_p$ is transitive (if $x >_p y$ and $y >_p z$, then $x >_p$ z);
2. indifference $\sim_p$ is transitive, reflexive ($x \sim_p x$ for all x), and symmetric ($x \sim_p y$ implies $y \sim_p x$);
3. exactly one of $x >_p y$, $y >_p x$, $x \sim_p y$ holds for each pair x and y; and
4. weak preference $\geq_p$ is transitive and complete (for a pair x and y, either $x \geq_p y$ or $y \geq_p x$).

Thus, an individual whose preferences can be represented by a weak order can rank all alternatives considered in a unique order. Further discussions of the properties of binary preference relations are presented in Fishburn (1970, Chapter 2) and Kreps (1990, Chapter 2).

Preference Representation Under Certainty

If strict preference $>_p$ on $\boldsymbol{X}$ is a weak order, then there exists a numeric representation of preference, a real-valued function v on $\boldsymbol{X}$ such that

$$x >_p y \text{ if and only if } v(x) > v(y),$$

for all x and y in $\boldsymbol{X}$ (Fishburn 1970). A preference representation function v under certainty is often called a value function (Keeney and Raiffa 1976). A value function is said to be order-preserving since the values $v(x)$, $v(y)$, ... ordered by $>$ are consistent with the preference order of x, y, ..., under $>_p$. Thus, any monotonic transformations of v will be order-preserving. As a result, the units of v have no particular meaning.

It may be desirable to consider a "strength of preference" notion that involves comparisons of preference differences between pairs of alternatives. To do so requires more restrictive preference assumptions, including that of a weak order over preferences between exchanges of pairs of alternatives (Krantz et al. 1971, Chapter 4). These axioms imply the existence of a real-valued function v on x such that, for all w, x, y, and z in $\boldsymbol{X}$, the difference in the strength of preference between w and x exceeds the difference between y and z if and only if

$$v(w) - v(x) > v(y) - v(z).$$

Furthermore, v is unique up to a positive linear transformation, i.e., if v' also satisfies the above difference inequality, then it must follow that $v'(x) = a\, v(x) + b$, where a (>0) and b are constants. This means that v provides an interval scale of measurement, such that v is often called a measurable value function to distinguish it from an order-preserving value function.

For multi-attribute decision problems, $\boldsymbol{X} = X_{1,}X_2 \ldots, X_n$, where n is the number of attributes and an element $x = (x_1, x_2, \ldots, x_n)$ in $\boldsymbol{X}$ represents an alternative. A multi-attribute value function can be written as $v(x_1, x_2, \ldots, x_n)$. Using some preference independence conditions, the multi-attribute value model can be simplified.

The subset $\boldsymbol{Y}$ of attributes in $\boldsymbol{X}$ is said to be preferentially independent of its complementary set $\bar{Y}$ if preferences for levels of these attributes $\boldsymbol{Y}$ do not depend on the fixed levels of the complementary attributes $\bar{Y}$. Attributes $X_1, X_2, \ldots, X_n$, are mutually preferentially independent if every subset of these attributes is preferentially independent of its complementary set.

A multi-attribute value function $v(x_1, x_2, \ldots, x_n)$ $n \geq 3$, has the following additive form

$$v(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} v_i(x_i), \tag{1}$$

where v_i is a value function over X_i if and only if the attributes are mutually preferentially independent (Keeney and Raiffa 1976; Krantz et al. 1971). When v is bounded, it may be more convenient to scale V such that each of the single-attribute value functions ranges from zero to one, leading to the following form of the additive value function:

$$v(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} w_i v_i(x_i), \tag{2}$$

where v and v_i are scaled from zero to one, and the w_i are positive scaling constants (usually called weights) summing to one. The assessment of models (1) and (2) are discussed in Keeney and Raiffa (1976, Chapter 3).

Dyer and Sarin (1979) proposed multi-attribute measurable value functions based on the concept of preference differences between alternatives that are much easier to assess than the additive form based on preferential independence. In addition to preferential independence, they considered some additional conditions that, loosely speaking, require that the decision maker's comparisons of preference differences between pairs of alternatives that differ in the levels of only a subset of the attributes do not depend on the fixed levels of the other attributes. These conditions allow the decomposition of a multi-attribute value model into additive and multiplicative forms. This development also provides a link between the additive value function and the multi-attribute utility model.

Preference Representation Under Risk

Perhaps the most significant contribution to the area of preference representation for risky options (i.e., lotteries or gambles) was the formalization of expected utility theory by von Neumann and Morgenstern (1947). This development has been refined by a number of researchers and is most commonly presented in terms of three basic axioms (Fishburn 1970).

Let P be a convex set of simple probability distributions or lotteries $\{X, Y, Z, \ldots\}$ on a nonempty set X of outcomes. (X, Y and Z will be used to refer to probability distributions and random variables interchangeably.) For lotteries X, Y, Z in P and all λ, $0 < \lambda < 1$, the expected utility axioms are:

A1. (*Ordering*) $> _p$ is a weak order;

A2. (*Independence*) If $X >_p Y$, then $\lambda X + (1 - \lambda) Z >_p \lambda Y + (1 - \lambda) Z$ for all Z in P;

A3. (*Continuity*) If $X >_p Y >_p Z$, then there exist some $0 < \alpha < 1$ and $0 < \beta < 1$ such that $\alpha X + (1 - \alpha) Z >_p Y >_p \beta X + (1 - \beta) Z$.

The von Neumann-Morgenstern expected utility theory asserts that the above axioms hold if and only if there exists a real-valued function u such that for all X, Y in P,

$$X >_p Y, \text{ if and only if } \mathrm{E}[\mathrm{u(X)}] > \mathrm{E}[\mathrm{u(Y)}],$$

where the expectation is taken over the probability distribution of a lottery. Moreover, such a u is unique up to a positive linear transformation.

The expected utility model can also be used to characterize an individual's risk attitude (Keeney and Raiffa 1976, Chapter 4). If an individual's utility function is concave, linear, or convex, then the individual is risk averse, risk neutral, or risk seeking, respectively. The von Neumann-Morgenstern theory of risky choice presumes that the probabilities of the outcomes of lotteries are provided to the decision maker. Savage (1954) extended the theory of risk choice to allow for the simultaneous development of subjective probabilities for outcomes and for a utility function u defined over those outcomes.

As a normative theory, the expected utility model has played a major role in the prescriptive analysis of decision problems. However, for descriptive purposes, the assumptions of this theory have been challenged by empirical studies (Kahneman and Tversky 1979). Some of these empirical studies demonstrate that subjects may choose alternatives that imply a violation of the independence axiom (A2). Prospect theory (Kahneman and Tversky 1979; Wakker 2010) attempts to explain these discrepancies. One implication of A2 is that the expected utility model is linear in probabilities. A number of contributions have been made by relaxing the independence axiom and developing some nonlinear utility models to accommodate actual decision behavior (Fishburn 1988).

For the case of multi-attribute decisions under risk, when $X = X_1 \times X_2 \times \ldots \times X_n$ in a von Neumann-Morgenstern utility model and the decision maker's preferences are consistent with some additional independence conditions, then $u(x_1, x_2, \ldots, x_n)$, can be decomposed into additive, multiplicative, and other well-structured forms that simplify assessment.

The attributes $X_1, X_2, \ldots, X_n$ are said to be additive independent if preferences over lotteries on $X_1, X_2, \ldots, X_n$ depend only on the marginal probabilities assigned to individual attribute levels, but not on the joint probabilities assigned to two or more attribute levels.

A multi-attribute utility function $u(x_1, x_2, \ldots, x_n)$, can be decomposed as

$$u(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} w_i u_i(x_i), \tag{3}$$

if and only if the additive independence condition holds, where u_i is a single-attribute function over X_i scaled from 0 to 1, and the w_i are positive scaling constants (or weights) summing to one. The additive model (3) has been widely used in practice.

If the decision maker's preferences are not consistent with the additive independence condition, a weaker independence condition that leads to a multiplicative preference representation may be satisfied.

An attribute X_i is said to be utility independent of its complementary attributes if preferences over lotteries with different levels of X_i do not depend on the fixed levels of the remaining attributes. Attributes $X_1, X_2, \ldots, X_n$ are mutually utility independent if all proper subsets of these attributes are utility independent of their complementary subsets.

A multi-attribute utility function $u(x_1, x_2, \ldots, x_n)$ can have the multiplicative form

$$1 + ku(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} [1 + kk_i u_i(x_i)], \qquad (4)$$

if and only if the attributes $X_1, X_2, \ldots, X_n$ are mutually utility independent, where u_i is a single-attribute function over X_i scaled from 0 to 1, the k_i are positive scaling constants, and k is an additional scaling constant. For approaches to the assessment of model (4) and other extensions of multi-attribute utility theory, see Keeney and Raiffa (1976).

The research of multi-attribute utility theory has been advanced from both theoretical and behavioral considerations. In particular, the effort of behavioral research tries to improve the descriptive power of multi-attribute utility models by incorporating psychological factors, such as aspiration level, goal and reference effect, and loss aversion (Tversky and Kahneman 1991). Various decision support systems have also been developed for multi-attribute decision making in the past decades, and applications of the theory and models have been expended to many new areas, including e-commence, public policy and environmental decisions, geographic information systems, and engineering (Dyer et al. 1992; Wallenius et al. 2008).

See

- ▶ Choice Theory
- ▶ Decision Analysis
- ▶ Multi-attribute Utility Theory
- ▶ Prospect Theory
- ▶ Utility Theory

References

Dyer, J. S., Fishburn, P. C., Steuer, R. E., Wallenius, J., & Zionts, S. (1992). Multiple criteria decision making, multiattribute utility theory: The next ten years. *Management Science, 38*, 645–654.

Dyer, J. S., & Sarin, R. K. (1979). Measurable multi-attribute value functions. *Operations Research, 27*, 810–822.

Fishburn, P. C. (1970). *Utility theory for decision making*. New York: Wiley.

Fishburn, P. C. (1988). *Nonlinear preference and utility theory*. Baltimore, MD: The Johns Hopkins University Press.

Kahneman, D. H., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–290.

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: Wiley.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundation of measurement*. San Diego, CA: Academic.

Kreps, D. M. (1990). *A course in microeconomics theory*. Princeton, NJ: Princeton University Press.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Tversky, A., & Kahneman, D. H. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics, 106*, 1039–1061.

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. New York: Cambridge University Press.

Wallenius, J., Fishburn, P. C., Zionts, S., Dyer, J. S., Steuer, R. E., & Deb, K. (2008). Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Management Science, 54*, 1336–1349.

Prescriptive Model

A model that attempts to describe the best or optimal solution of a man/machine system. For a decision problem, such a model is used as an aid in selecting the best alternative solution.

See

- ▶ Decision Problem
- ▶ Descriptive Model
- ▶ Mathematical Model
- ▶ Normative Model

Prices

In the simplex method, for a nonbasic variable x_j, the price is defined as $d_j = c_j - z_j$ or $d_j = z_j - c_j$, where c_j is the variable's original cost coefficient and $z_j = \boldsymbol{\pi}\mathbf{A}_j$, with $\mathbf{A}_j$ the variable's original column of coefficients and $\boldsymbol{\pi}$ the multiplier (pricing) vector of the current basis. The d_j is termed the reduced or relative cost. It is the difference between the direct cost c_j and indirect cost z_j. The d_j indicates how much the objective function would change per unit change in the value of x_j. The d_j for the variables in the basic feasible solution are equal to zero.

See

- ▶ Devex Pricing
- ▶ Opportunity Cost
- ▶ Simplex Method (Algorithm)

Pricing Multipliers

- ▶ Multiplier Vector

Pricing Out

In the simplex method, the calculation of the prices associated with the current basic solution.

See

- ▶ Prices
- ▶ Simplex Method (Algorithm)

Pricing Vector

- ▶ Multiplier Vector
- ▶ Prices
- ▶ Simplex Method (Algorithm)

Prim's Algorithm

A procedure for finding a minimum spanning tree in a network. The method starts from any node and connects it to the node nearest to it. Then, for those nodes that are now connected, the unconnected node that is closest to one of the nodes in the connected set is found and connected to these closest nodes. The process continues until all nodes are connected. Ties are broken arbitrarily.

See

- ▶ Greedy Algorithm
- ▶ Kruskal's Algorithm
- ▶ Minimum Spanning Tree Problem

Primal Problem

The primal problem is usually taken to be the original linear-programming problem under investigation.

See

- ▶ Dual Linear-Programming Problem

Primal-Dual Algorithm

An adaptation of the simplex method that starts with a solution to the dual problem and systematically solves a restricted portion of the primal problem while improving the solution to the dual. At each step, a new restricted primal is defined and the process continues until solutions to the original primal and dual problems are obtained.

See

- ▶ Simplex Method (Algorithm)

Primal-Dual Linear-Programming Problems

▶ Dual Linear-Programming Problem
▶ Linear Programming

Principle of Optimality

Condition that Richard Bellman derived for dynamic programming: "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision." (Bellman 1957, Chap. III.3)

See

▶ Bellman Optimality Equation
▶ Dynamic Programming

References

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Prisoner's Dilemma

A two-person game where neither player knows the other's play (action or decision) a priori. Imagine a situation where two criminals are isolated from each other and the police interrogator offers each the following deal: if the prisoner confesses and the confession leads to the conviction of the other prisoner, he goes free and the other prisoner gets 10 years in prison. However, if both confess, they each get 5 years. If neither confesses, there is enough evidence to convict both on a lesser offense and they both get one year. If there is no trust, then both will confess, whereas if there is complete trust, neither will. Since complete trust is rare, when the game is played one time, players almost always defect. When the game is played repeatedly and there is a chance for a long-term reward, wary cooperation with a willingness to punish defection is the best strategy. This game illustrates many social and business contracts and is important for understanding group behavior, both cheating and cooperation. It has also been used in studying political and military strategies.

See

▶ Game Theory

References

Poundstone, W. (1992). *Prisoner's dilemma*. New York: Doubleday.

Probabilistic Algorithm

An algorithm that employs probabilistic elements (as opposed to a deterministic algorithm).

See

▶ Genetic Algorithms
▶ Randomized Algorithm

Probabilistic Programming

A mathematical programming problem in which some or all of the data are random variables.

See

▶ Chance-Constrained Programming
▶ Stochastic Programming

Probability Density Function (PDF)

When the derivative $f(x)$ of a cumulative probability distribution function $F(x)$ exists, it is called the density or probability density function.

Probability Distribution

Term used (loosely) to refer to a function describing the probabilistic behavior of a random variable; could refer to the probability measure, the cumulative distribution function (CDF), the probability mass function (PMF) for discrete random variables, or the probability density function (PDF) for continuous-valued random variables.

Probability Generating Function

For a non-negative integer-valued random variable X with probability mass function $p_j = \Pr\{X = j\}$, the probability generating function (often just called the generating function, and known in other fields as the z-transform) is given by $P(z) = E[z^X] = \sum_{j=0}^{\infty} z^j p_j$. The definition can be extended to the setting where X can take all integer values (i.e., including all negative values).

Probability Integral Transformation Method

One of the primary methods for generating random variates for Monte Carlo or discrete-event simulation, using the cumulative distribution function (CDF) commonly known as the inverse transform method.

See

- ▶ Inverse Transform Method
- ▶ Random Number Generators
- ▶ Random Variates
- ▶ Simulation of Stochastic Discrete-Event Systems

Probability Mass Function (PMF)

Function giving the probability of taking on each of the possible discrete values.

Problem Solving

The process of deciding on actions aimed at achieving a goal. Initially, the goal is defined to represent a solution to a problem. During the reasoning process, subgoals are formed, and problem solving becomes recursive.

See

- ▶ Artificial Intelligence
- ▶ Decision Analysis
- ▶ Decision Making and Decision Analysis
- ▶ Decision Support Systems (DSS)
- ▶ Expert Systems

Problem Structuring Methods

Jonathan Rosenhead
The London School of Economics and Political Science, London, UK

Introduction

Problem structuring methods (PSMs) are a broad group of model-based problem handling approaches whose purpose is to assist in the structuring of problems rather than directly to derive a solution. They are participative and interactive in character, and normally operate with groups rather than individual clients. In principle they offer OR/MS access to a range of problem situations for which more classical OR techniques have limited applicability. The most widely adopted of these methods are Soft Systems Methodology, the Strategic Choice Approach, and Strategic Options Development and Analysis (SODA).

PSMs developed out of, or at least intertwined with, a critique of the restricted scope of traditional OR techniques. From the 1970s there developed an active debate over claims for the objectivity of OR/MS models, and about the limitations imposed on OR/MS practice by its concentration on well-defined problems. Significant critical contributions were made by

Rittel and Webber (1973), Ackoff (1979), Checkland (1981), Rosenhead and Thunhurst (1982), Eden (1982), Rosenhead (1986), Jackson (1987), Flood and Jackson (1991), Mingers (1992). The general thrust was that standard OR techniques assume that relevant factors, constraints, and objective function are both established in advance and consensual; commonly the function of the technique is to determine an optimal setting of the controllable variables. Consistently with this, standard formulations of OR methodology were seen to assume a single uncontested representation of the problematic situation under consideration.

Critics have recognized that OR's practice has been considerably more diverse than this, and in particular is far from dominated by considerations of optimality; however, the available tools were held to offer little appropriate assistance outside this area. The methodological framework on offer was equally seen as giving scant guidance to analysts confronting less well-behaved circumstances. There are situations in which intangibles, uncertainty, and value diversity as well as complexity are crucial presences. Skilled operational researchers have been able to make progress in such situations, but only by using tacit skills which are not part of the OR/MS canon. Yet the more socially important the decision situation, the more likely it is that such features will come to dominate.

Out of this critique of the shortcomings of traditional OR/MS a family of alternative methods was developed, with both common features and also differences of focus. When their similarities were recognized the label used to describe them as a group was Problem Structuring Methods (though other names such as Soft OR are also in currency). The over-arching emphasis which the methods share is on helping groups of decision-makers to identify what problem they could usefully work on together, and to assist them in making progress with that task. There is no assumption that the decision-makers share a common perspective, so that they are perhaps more accurately described as stakeholders. Nor are these methods to any significant degree quantitative. This is because the approaches are all based on the participation of those who have the problem. If mathematics were to be the language of the discourse, some (perhaps many) of the participants would be disempowered, or at least prevented from enunciating perceptions important to them which could not be expressed in that format, or only by a distortion which changed their content.

Each of the methods within the PSM family consists of a number of technical procedures linked together through social processes. i.e., unlike the algorithmic approaches that have tended to dominate OR/MS, the consultant does not identify and then input some starting conditions from which the 'answer' will be produced without further human intervention. What happens is that at various points the groups discuss the implications of the analysis to date, and on that basis (and aided by a facilitator) decide how to proceed further, or maybe whether enough progress has been made that the stakeholders can proceed without further analytic assistance. For clarity one should perhaps describe PSMs as 'methodologies' rather than 'methods', taking a methodology as an assembly of technical and process elements.

In short these methods bear very little resemblance to those developed within traditional OR/MS. The one key unifying element is the central use of cause-effect models. Each of them uses formal models to represent the problematic situation perceived by the decision-making group, in order to summarise, coordinate and advance their understanding of the situation they confront. The types of model used are specific to each method, but none of them are 'computable'. Indeed quantification has little if any role in any of them. The concepts that are in play are more usually verbal, and the operations on them are mostly performed by the group, who through discussion transform the models based on their changing understanding. The outcomes of a successful application of a PSM will be a group of decision-makers confident enough to take action; a group of decision-makers who have gained a deeper insight into their problem area; and a group of decision-makers whose shared experience has led to improved relations with each other.

Types of Problem

Before going on to outline the PSM field, it should be helpful to address the apparent paradox of two very different types of methodology, sometimes called 'hard' (i.e., traditional) and 'soft' (PSMs), each addressing problems of complexity in an analytic manner. One simple explanation lies in the quite wide recognition of two substantially different types of problem situation. Rittel and Webber's (1973) characterization of them as tame vs. wicked has achieved wide currency, as has Schon's (1987) extended metaphor of problems of the swamp contrasted with those of the high ground. Tame problems (on the high ground) have precise, unproblematic formulations permitting powerful analyses of great technical sophistication. Wicked problems (in the swamp) have multiple stakeholders, intangible objectives, key uncertainties, contested or doubtful formulations, etc. In the latter there is no unified representation of the issue or issues that can be established *ahead* of analysis. Rather, a representation or representations of the problematic situation which participants find helpful may be a major product of the analysis.

It follows from this diagnosis that methods that are designed to be effective in handling tame problems are likely to be largely irrelevant for wicked ones. (And vice versa of course.) For the latter type of problem situation, methods that assist argumentation, promote negotiation or generate mutual understanding are needed, rather than those that reliably and efficiently identify an optimum. Methods that can only start once there is an agreed problem (but have no methods for reaching that agreement) are liable to ignore or dismiss alternative perspectives and their contrary formulations.

The much remarked difficulty which OR/MS encountered from the late 1960s in securing access to more strategic levels of decision-making may be attributed at least in part to this factor. As Schon observed, problems of major social importance are commonly located in "swamp" conditions. Attempts to address these "messes" using techniques and methodology developed for handling well-structured problems constituted inappropriate technology transfer. Where solutions based on these methods were adopted they were vulnerable to being savaged in practice by the 'wicked' parts of the problem situation that had been excluded. More commonly however such representations were recognised as an overly thin representation of the rich and complex world that managers and decision-makers inhabit – with the result that OR/MS was confined to the tame (less strategic, more repetitive, operational) aspects of organisational life.

Characteristics of Alternative Methods

Problem structuring methods constitute a family of approaches offering appropriate support to decision-making under these less pacified circumstances. They were developed separately by individual innovators or teams of innovators, and each emphasizes or is organized around particular aspects of the wicked problem environment. Indeed each had been independently developed before a recognition arose of their family resemblance. (Subsequent to that recognition, however, many of the principal originators entered into a constructive dialogue with each other, in which a certain amount of mutual borrowing of particular elements took place.: For example, distinctive post-it 'Ovals', originated for use within SODA, became widely used by other methods.) In other words the new methods grew out of practice. However their similarities are by no means coincidental. Many leading developers of PSMs had been active participants in the critique of traditional methods, and their innovations were designed to remedy particular inadequacies of the conventional repertoire in handling wicked problems. So at that fundamental level there was a common theoretical base.

PSM methods have differing rationales, purposes, technical apparatus, etc. Some of these distinctive attributes will be indicated below. However it will be useful, first, to identify the features which they hold in common.

Rosenhead (1989) has provided one formulation, based on inverting the characteristics of the conventional OR/MS paradigm.

PSMs

- Seek solutions which satisfice on separate dimensions (rather than trade-off onto a single dimension to facilitate optimization);
- Integrate hard and soft data with social judgments (reducing data greed with its problems of quality and distortion);
- Produce transparent models which clarify any conflicts (rather than basing a scientific depoliticization on an assumed consensus);
- Treat people as subjects actively engaged in the decision-making process (rather than as passive objects to be modelled or disregarded);
- Facilitate planning from the bottom-up (and not as a process driven by the abstract objectives of a hierarchically located decision-maker); and

- Accept that some uncertainty is irreducible and aim to preserve options (rather than base current and future decisions on a notionally certain future).

The methods clearly assume a decision-making quite different from that of conventional OR/MS applications, and this environment places particular requirements on the interface with the client group. Where consensual values cannot be assumed, there will be a need to achieve agreement among a range of stakeholders representing different interests and/or holding different perspectives. It follows from this that a PSM should be able to accommodate multiple alternative perspectives, often in a group situation in which holders of those viewpoints are present and participating. From this it follows that for a method to be helpful it must operate iteratively and interactively; as participants internalize and adjust to each others' contributions, new formulations of the problematic situation will emerge which in turn feed new modelling and structuring activity. And since participants have different though overlapping organizational agendas, and also because of the prevalence of uncertainty, any resulting consensus on action is likely to constitute a partial rather than a comprehensive solution to the problems present within the situation under discussion.

These social requirements on a PSM have implications for the technical repertoire that it can deploy. Its handling of complexity must not obstruct lay participation — which points to graphical (rather than, for example, algebraic) representations. The existence of multiple perspectives invalidates the search for an optimum; the need is rather for systematic exploration of the solution space. To elicit meaningful judgments from lay participants, abstract continuous variables need to be eschewed in favour of discrete concrete alternatives that can be compared. And, given the need to avoid illusions of precision when confronting uncertainties, possibilities will be more helpful than probabilities, and alternative scenarios will enrich discussion that forecasts might close down.

These outline specifications for a more appropriate decision-aiding technology eliminate much of the scope for advanced mathematics, probability theory, complex algorithms. They identify, rather, an alternative approach employing representation of relationships, symbolic manipulation, and limited quantification within a systematic framework. These are decidedly low-tech methods: some of them have no software support, and even those that do can be operated in manual mode. The lack of mathematics should not however be taken for lack of rigour. These are methods with their own rigour, which is qualitative in nature.

The Methods

There is no definitive list of problem structuring methods. However to give identity to the field it is appropriate to provide some demarcation criteria.

PSMs

- Can be distinguished from traditional OR methods by the six criteria listed in the previous section.
- Can be distinguished from non-OR modes of working with groups, such as Organizational Development, by the core element of an explicit modelling of cause-effect relationships.
- Can be demarcated from other OR approaches which purport to tackle messy, ambitious problems (e.g., the Analytic Hierarchy Process) by PSMs' transparency of method, restricted mathematization, and focus on supporting judgment rather than representing it.

These limits are imprecise and arguable; and there is scope for approaches developed for other or broader purposes (e.g., spreadsheet models) to be used in a similar spirit. Ackoff's Interactive Planning is close in both spirit and intent (see Ackoff 1999) but nevertheless has never been regarded as falling within PSMs. (Rather than changing this de facto if not de jure circumstance, it will not be discussed further.) Methods that have some degree of similarity to PSMs but also significant differences are (for coherence) best regarded as falling outside the category. These include multi-criteria decision methods, outranking methods such as PROMETHEE and ELECTRE, decision conferencing, scenario planning, system dynamics (in some of its versions) and Viable System Diagnosis. Other parts of the PSM perimeter are bordered by the focus group approach, and by Rapid Rural Appraisal and other participative third world development approaches (for which see Rosenhead and Mingers 2001, pp. 345-7).

A brief introduction to the better established PSMs follows (Rosenhead and Mingers 2001):

Strategic Options Development and Analysis (SODA)

This method is described fully in Eden and Ackermann (1998). It is a general purpose problem identification method that uses cognitive mapping as a modelling device. The concepts that individuals use to make sense of their problematic situation, and the causal links thought to exist between those concepts, are elicited in individual interviews and recorded in map form. The maps drawn from separate interviews with stakeholders are subsequently merged into a single 'strategic map' through pinning together concepts common to more than one of them. The strategic map, commonly structured into clusters, provides the framework for discussion in a workshop of the group of map 'owners', at which a facilitator uses the map to guide participants towards commitment to a portfolio of actions. An alternative and more rapid version known as the Oval Mapping Technique operates in workshop mode throughout, and can in principle achieve results in a 1 day session. The participants commit their concepts to 'Ovals' (specially designed PostIt notes), which the facilitator with the participation of workshop members organises into an agreed structure. This then serves as the strategic map for the discussion that follows.

Soft Systems Methodology (SSM)

Soft Systems Methodology is a general method for system design or redesign, which aims to generate debate about alternative system modifications. It adopts a systems theoretic framework for exploring the nature of problem situations, and how purposeful action to change them might be agreed when there are different perceptions of the situation based on contrasting world views. A systematic exploration of the world views of stakeholders leads to the generation of definitions of alternative systems, the investigation of which is expected to be of interest from at least one of those world views. Each of these abstract 'root definitions' is expanded into the component activities which would be necessary for it to operate successfully. This generates a range of contrasting alternatives for the modification of the system, which are used to generate debate about which changes are both culturally feasible and systemically desirable. Full descriptions of the method are available in Checkland (1981, 2006, 1990).

Strategic Choice Approach (SCA)

Strategic Choice is a planning approach centred on the management of uncertainty and commitment in strategic situations. Typically a Strategic Choice engagement takes place entirely in workshop format, with no backroom work by the consultants. There are four modes of analysis:

- Shaping – in which different areas for choice are elicited from workshop members. A subset of these is selected as a problem focus by reference to their urgency, importance and inter-connectedness
- Designing – here the options for action for each of the decision areas within the problem focus are identified, as well as any incompatibilities between option selections in different decision areas. The feasible decision schemes (consisting of one option choice within each decision area) are derived
- Comparing – criteria for choice, often non-quantitative, are agreed by the group. These are used first in satisficing mode to establish a working shortlist of schemes; pairwise comparisons of shortlisted schemes are made, establishing on each criterion a range of relative advantage between the two schemes. This may be repeated for different pairs. Commonly significant uncertainties are revealed by this process. Other uncertainties will usually have been identified in previous modes
- Choosing – bearing in mind the surfaced uncertainties, a 'progress package' is agreed consisting of partial commitments to be made at this stage, explorations to be launched to reduce key uncertainties, contingency plans, and a timetable for later choices.

Facilitators assist with the deployment of the transparent tools available within the method, and in guiding the, possibly recursive, switching between modes. A detailed account of the method is available in Friend and Hickling (2004).

Robustness Analysis

Robustness Analysis is another approach for use where uncertainty is an important issue. It focuses on one specific strategy for managing that uncertainty - that of maintaining useful flexibility. The focus of the approach is on initial commitments rather than on future plans for the system. The flexibility of an initial commitment relates to its compatibility with a range of acceptable or desirable future states of the system. It is this flexibility left by an initial commitment that is operationalised as a decision-making criterion by the concept of the robustness. This is defined as a ratio where the denominator is the number of states whose performance at the planning horizon is ‘good enough’; the numerator is the number of those states which would remain accessible if the commitment under consideration were to be made. Robustness analysis can be conducted with either a single or multiple futures employed to estimate system performance; and it can be used in conventional or interactive mode. In the latter, participants and analysts assess both the compatibility of initial commitments with possible future configurations of the system, and the performance of each configuration in feasible future environments. This enables them to compare the flexibility maintained by alternative initial commitments. It is in this latter mode that Robustness Analysis qualifies as a PSM, though even when used in non-participatory mode it maintains an accessible transparency. For more detail, see Rosenhead and Mingers (2001).

Drama theory

Drama Theory draws on two earlier approaches, metagames and hypergames. It is an interactive method of analysing co-operation and conflict among multiple actors. A model is built from perceptions of the options available to the various actors, and how they are rated. Drama theory looks for the ‘dilemmas’ presented to the actors within this model of the situation. Each dilemma is a change point, tending to cause an actor to feel specific emotions and to produce rational arguments by which the model itself is redefined. When and only when such successive redefinitions have eliminated all dilemmas is the actors’ joint problem fully resolved. Analysts commonly work with one of the parties, helping it to be more effective in the rational-emotional process of dramatic resolution. For more detail, see Howard (1999).

Applications of PSMs

As can be inferred from their remit to structure wicked problems, the problem situations to which PSMs have been applied have a wide variety. A good source for practical applications of the SCA is Chapter 13 of Friend and Hickling 2004, pp. 298-360. An overview of applications across the range of PSMs is provided by Mingers and Rosenhead (2004), which is the review article for a special issue of the *European Journal of Operational Research* on applications of PSMs (Vidal 2004).

A diverse record of successful applications is an indicator of wide relevance, but a disadvantage when it comes to providing a coherent summary. A literature survey covering the period up to 1998 (summarized in Mingers and Rosenhead 2004) categorises 51 reported applications under the headings general organizational/information systems/technology, resources, planning/health services/general research. Two comments seem appropriate: (i) it is plausible to assume that reported cases are the tip of the iceberg; and (ii) 1998 was relatively early in the development of interest in PSMs.

The categories supplied in the previous paragraph are so broad as to give little flavour of the reality of PSM practice. To provide that, some short summaries of projects using PSMs that are described in Mingers and Rosenhead (2004) may be of assistance

- Organisational restructuring at Shell. SSM used to provide the basis of a reconfiguration of a central department of Shell International, in a series of workshops with senior managers
- Models to support a claim for damages. SODA (as well as System Dynamics) used to support a legal case by the Canadian-based multinational Bombardier against Trans Manche Link, for damages resulting from delays in processing designs for the Channel Tunnel shuttle wagons
- Supporting a tenants cooperative. This was an engagement over several years to help a cooperative of residents of an ex-mining village to manage their own housing. Elements of various PSMs, as well as other methods (e.g., spreadsheet financial models) were used to support strategic decisions, and help the cooperative gain confidence
- IT strategy for a supermarket chain. This study reported to the joint chief executives of the leading British supermarket chain Sainsbury’s, and worked with a 16-strong senior management task force. SODA, SSM and SCA were all used at different stages, to identify IT systems that would support business objectives

- Planning for a street festival. The largest European street festival (Notting Hill Carnival) was a victim of its own success, with issues of security, congestion, cultural integrity etc. Working with representatives of the carnivalists, local government, transport and emergency services, and arts organisations, SSM and SCA were used to devise escape strategies
- National level planning in Venezuela. A version of SCA has been used at various levels of the state service in Venezuela, up to and including the Cabinet, to agree on strategic decisions in a range of areas
- Local pediatric care strategy. Health care managers and specialists in an Inner London area with some 500,000 population needed to reduce the number of inpatient paediatric care units. SCA was used in a series of workshops to produce agreement between representatives of all stakeholders on (i) how many units should remain (ii) where they should be; and (iii) what consequential changes were needed to other aspects of the health service.

This list indicates the reach of these methods, from grass-roots community groups through senior corporate management issues to the highest levels of national government. The content of many if not all of the projects would have rendered them inaccessible to conventional OR.

Using PSMs

Working with Clients

PSM practitioners have to be able to manage not only the complexity of substantive subject matter but also the dynamics of interaction among workshop participants. The dual roles of analyst and of facilitator of group process place heavy demands on the consultant, who is called upon to deploy a wider range of skills than in conventional operational research practice. When operating as facilitator she has the responsibilities of ensuring that all voices are heard (not suppressed by psychological or hierarchical effects); that apparent agreement is not based on mutual misunderstanding of key terms; and that the precious (and usually expensive) opportunity presented by the gathering of key stakeholders is exploited in a timely and effective manner. (This experience is hard to simulate ‘off line’, and training should, if possible, include at least a brief experience of practical apprenticeship.) It is useful to have two facilitators with differentiated roles. One of them is likely to be heavily engaged, at times leading the discussion, at others concentrating acutely on the content of the discourse and also on the interpersonal issues that it reveals. The second facilitator can be principally involved with keeping a record, perhaps by direct computer input, of the evolving model. But he will also be able to intervene with insights that his colleague might otherwise miss through following the scent too closely.

PSMs are based on the working assumption that the client is not a sole decision-maker but a client system. Organisational politics is thus an integral aspect of project process, to which the consultants must be sensitive if they are not to be derailed. In order to achieve an effective process and worthwhile outcome it is important that all relevant stakeholders are represented. This requirement may bump up against numerical constraints – most practitioners cite a group size in the range 6–10 as desirable, and 12 should be the absolute maximum for a coherent group conversation to take place. There may be pressures to add people beyond this number for reasons of organisational politics, or to exclude certain clearly relevant stakeholders. These issues of the design of the group are ones that the consultant must address.

To guide the workshop with the consent and indeed respect of the group, the consultant must be, and be seen to be, disinterested – that is, not operating on behalf of any sectional interest. Where political tensions are active, this can require both sensitivity and agility from the facilitator. In inter-organisational working (for which the multi-perspective approach of PSMs makes them particularly appropriate), the question of access to the problem domain potentially acquires an additional twist. Initial contacts with one of the organisational actors will be necessary to gain entry to the problem forum - but that entry route may itself occasion doubts among other stakeholders as to the impartiality of the facilitation that follows.

Selecting Methods

There is no established process for the selection of method or methods to use in a particular engagement. This is often done on an intuitive basis – where uncertainties are seen as particularly salient in the problematic situation, Strategic Choice or Robustness

Analysis are plausible candidate methods; an evident conflict situation may suggest drama theory; and so on. There is of course also a choice to be made between using a traditional method or a PSM of any kind.

The most widely cited and discussed framework for this higher-level choice is due to Jackson and Keys (1984). Their 'system of systems methodologies' proposed two dimensions on which to describe the context of a problem. These were the degree of agreement among participants – which can be unitary (consensus), pluralistic (several viewpoints but agreement possible), or coercive (disagreements resolved through exercise of power); and the nature of the problem - simple or complex. This yielded six cells into which OR/systems methods were placed. For example, traditional hard OR was most suitable for simple–unitary contexts, System Dynamics for complex–unitary contexts, and Problem Structuring Methods for complex–pluralistic contexts.

However the criticism has been advanced that this framework makes the (unwarranted) assumption that the nature of the problem context can unerringly be identified in advance. Commonly, however, this will not be the case in the messy situations that PSMs are appropriate for. It may well be that only *after* the investigation is underway will the view to be taken of the problem context become clear. Furthermore, since the use of PSMs is a form of organised finding out, it is quite possible that this process will change the initial understanding of the problem context. For example, what was initially perceived by the relevant actors to be pluralist in character may, as a result of the intervention be reperceived as falling elsewhere on the spectrum of degree of agreement.

Mixing Methods

Another feature undermining the simplicity of the Jackson and Keys scheme is the fact that many PSMs consist of a loosely articulated set of processes (part technical, part social), with considerable freedom to switch phase or to recycle. They therefore lend themselves to creative re-assembly, in which different methods or parts of different methods are used in conjunction. Before theoretical discussion of this potential took off in the 1990's it was already a de facto reality in practice. The most high profile of many applications was the Sainsbury's case study (Ormerod 1996) already mentioned above, in which SODA, SSM and SCA were employed on a single engagement. In fact several of the cases summarised in the Applications section of this article involved the use of parts or wholes of PSMs in combination, or indeed the joint use of a PSM with a more conventional OR technique.

This ongoing practice was systematised and given a theoretical base by multimethodology (Mingers and Gill 1997). This advocates seeking to combine together a range of methods, perhaps across the hard/soft divide, in order to deal effectively and appropriately with the qualitatively different analytic challenges which a single problem situation may pose. Based on the work of Habermas (1984, 1987), any real-world problem situation can be seen as a complex mix of the material, the social, and the personal. Different methods are appropriate for analysing and making progress in these different strata. Thus material or physical characteristics can be modelled using traditional OR techniques, but social conventions, politics and power, and personal beliefs and values need quite different, qualitative approaches.

Any practical project goes through several stages - understanding and appreciating the situation, analysing information, assessing different options, and acting to bring about change. Moving from one phase to another offers an opportunity to transfer, based on the understanding achieved up to that point, to a different level (say from the material to the social) and to a corresponding type of analysis. The appropriate use of varied methods allows the project to evolve creatively, rather than pursuing the methodology adopted at its start, regardless of the understanding which is progressively developed. These are complementary arguments for combining together different PSMs, and indeed PSMs with other methods. Multimethodology facilitates a more varied palette, to match the developing richness of problem understanding.

Software and Other Technology

Several established PSMs have associated software: examples include STRAD (for Strategic Choice) and Decision Explorer (for SODA). These packages perform a variety of functions. They may display and re-organize concepts and their inter-relationships; identify a feasible range of options for action; elicit preferences using paired comparisons; compute simple quantitative attributes of options derived from the current problem structure, and so on. They may also perform a variety of roles in the project, from technical assistance to the facilitator between group sessions, through enabling individual participants to pursue solo investigations, to the provision of an online Group Decision Support System. The use of the software during group sessions undoubtedly has an effect on group dynamics, focusing attention and giving a degree of control to whoever is in charge of inputting data or of changing the visual display. For this reason some leading practitioners prefer not to employ computers during the actual workshop sessions. In SODA, however, the computer display of sections of strategic maps is used deliberately in order to influence the group conversation. The computer model (i.e., map) is deployed as a 'facilitative device', so that group members will more easily accept and absorb concepts that are new to them. A concept that is advanced by another group member might provoke resistance – but one which whose presentation is neutrally framed by the computer may be easier to accommodate to.

The distinctive technology of PSMs is low- rather than high-tech. Ongoing models and other notes on deliberations are recoded on A1 flipchart sheets on the meeting room walls. Oval 'postit' notes are used to capture concepts in a way which facilitates re-structuring of model relationships during the session. At the end of a workshop it is normal for these traces to be photographed, and then emailed to participants. This visual record is a vivid reminder not only of the outcome of a workshop, but also of the process by which it was reached.

Implementation

A PSM workshop should leave time at the end for the group to agree an implementation strategy. If it is an intermediate workshop with others to follow, this process will constitute the allocation of responsibilities to group members (including the facilitators) to pursue clarification or uncertainty reduction activities that have been revealed as advantageous, so that the following meeting can take off from an improved position. With some PSMs the intervening work may consist of model development by the consultants – e.g., producing revised SSM 'root definitions'; in SODA reflecting the discussion in redrawn maps; and in SCA carrying out explorations to reduce relevant uncertainties. At what is expected to be a final workshop, where some conclusions have been agreed, the implementation strategy needs to be articulated and bought into by the key players. This will require a thorough discussion to identify the tasks (including for example a dissemination strategy) necessary for sustainable action to take place, and to specify responsibilities for these.

The experience within a PSM workshop, when it is working well, is frequently intense and the sense of release and satisfaction when a breakthrough is made can be palpable. Negotiated accommodations arrived at in this way can be creative escapes from apparently irresolvable tangles. However this almost cathartic experience is not transferable to non-participants. Generally only a part of the client system will be present at the workshop, and those not present may be reluctant to take its outputs on trust. Indeed it is more likely than not that those who can actually set the wheels in motion have not been members of the workshop. A report in conventional form which presents the case for the decisions arrived at in linear fashion the may be needed. For work within a single hierarchically structured organisation, top-down authority may carry the outputs of a PSM-based process towards implementation. In the case of inter-organisational work the situation is more complex, and the generation of acceptance among the various organisational constituencies can be problematic. It is clearly advisable for these problems of multiple acceptance to be discussed by the group, and to inform the implementation strategy.

Concluding Remarks

The progress of Problem Structuring Methods - in development and sophistication of methods, in applications, and in geographic spread - since they were recognised as a category with strong family resemblances has been fairly uninterrupted. There is one exception: the United States. The development of PSMs has been virtually ignored by the US OR/MS community. This was pointed out in an unprecedented letter to *ORMS Today* (Ackerman et al. 2009) signed by 45 academics from 11 countries and four continents. They cited as a strong contributory factor the systematic exclusion of papers on this topic from

US-based academic journals. An article in the same issue (Mingers 2009) explored the phenomenon in greater depth. For a further analysis of the difference in treatment of PSMs between the U.S. and the U.K., see Paucar-Caceres (2011).

In much of the rest of the world, PSMs have effected a breakout from the well developed but relatively confined arena of technocratic solutions to consensually defined problems occupied by OR's traditional methods. This outward movement has brought decision-support modelling in touch with a range of other methods and practices designed to help groups make progress with their problems. It has been suggested elsewhere (Rosenhead and Mingers 2001) that large group methods, development planning methods and community operational research are among the areas from which PSMs can learn, and to which PSMs can contribute.

The presence of Community OR (Midgley and Ochoa-Arias 2004) in this list is due to its natural fit with PSMs. Community OR is an analytic practice aimed at extending the customers of OR to include disadvantaged and non-hierarchical groups. With few resources, many of traditional OR's resource allocation tools are irrelevant. Furthermore the weak are perhaps disproportionately confronted with 'wicked', less well-structured problems; and the bottom-up nature of the PSM approach seems appropriate for the defined clientele. Its transparent modelling approach and group orientation does not present as many obstacles to engagement as would traditional OR's more mathematical approaches. No doubt these are among the reasons for the relatively high penetration of PSMs in this area.

There is now a substantial record of achievement for PSMs. There have been a wide variety of different types of use, both in context and in content. Surveys have shown there to be a good measure of user satisfaction. And there is an exciting range of possible further developments which appear to be reachable from the base that has already been achieved.

See

- ▶ Community OR
- ▶ Practice of Operations Research and Management Science
- ▶ Robustness Analysis
- ▶ Soft Systems Methodology
- ▶ Strategic Choice Approach (SCA)
- ▶ Strategic Options Development and Analysis (SODA)
- ▶ System Dynamics
- ▶ Wicked Problems

References

Ackerman, F., Bawden, R., et al. (2009). The case for soft or letter to the editor. *ORMS Today, 36*, 20–21.

Ackoff, R. L. (1979). The future of operational research is past. *Journal of the Operational Research Society, 30*, 93–104.

Ackoff, R. L. (1981). The art and science of mess management. *Interfaces, 11*, 20–26.

Ackoff, R. L. (1999). *Re-creating the corporation: A design of organization for the 21st century*. New York: Oxford University Press.

Checkland, P. B. (1981). *Systems thinking systems practice*. Chichester: Wiley.

Checkland, P., & Poulter, J. (2006). *Learning for action: A short definitive account of soft systems methodology, and its use for practitioners teachers and students*. Chichester: Wiley.

Checkland, P., & Scholes, J. (1990). *Soft systems methodology in practice*. Chichester: Wiley.

Eden, C. (1982). Problem construction and the influence of OR. *Interfaces, 12*, 50–60.

Eden, C., & Ackermann, F. (1998). *Making strategy: The journey of strategic management*. London: Sage.

Flood, R. L., & Jackson, M. C. (1991). *Creative problem solving: Total systems intervention*. Chichester: Wiley.

Friend, J., & Hickling, A. (2004). *Planning under pressure: The strategic choice approach* (3rd ed.). Oxford: Elsevier.

Habermas, J. (1984). *The theory of communicative action. Vol. 1: Reason and the rationalization of society*. London: Heinemann.

Habermas, J. (1987). *The theory of communicative action. Vol. 2: Lifeworld and system: A critique of functionalist reason*. London: Heinemann.

Howard, N. (1999). *Confrontation analysis: How to win operations other than war, CCRP*. Washington, DC: Department of Defense.

Jackson, M. C. (1987). Present positions and future prospects in management science. *Omega, 15*, 455–466.

Jackson, M. C., & Keys, P. (1984). Towards a system of systems methodologies. *Journal of the Operational Research Society, 35*, 473–486.

Midgley, G., & Ochoa-Arias, A. E. (Eds.). (2004). *Community operational research: OR and systems thinking for community development*. New York: Kluwer.

Mingers, J. (1992). Recent developments in critical management science. *Journal of the Operational Research Society, 43*, 1–10.

Mingers, J. (2009). Taming hard problems with soft O.R. – ‘Soft’ methodologies tackle messy problems that traditional O.R. can’t touch, so why isn’t it promoted in the U.S.? *ORMS Today, 36*, 48–53.

Mingers, J., & Gill, A. (Eds.). (1997). *Multimethodology: The theory and practice of combining management science methodologies*. Chichester: Wiley.

Mingers, J., & Rosenhead, J. (2004). Problem structuring methods in action. *European Journal Operational Research, 152*, 530–554.

Ormerod, R. J. (1996). Information systems strategy development at sainsbury’s supermarkets using “Soft” ORC. *Interfaces, 26*, 102–130.

Paucar-Caceres, A. (2011). The development of management sciences/operational research discourses: surveying the trends in the US and UK. *Journal of the Operational Research Society, 62*, 1452–1470.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Science, 4*, 155–169.

Rosenhead, J. (1986). Custom and practice. *Journal of the Operational Research Society, 37*, 335–343.

Rosenhead, J. (Ed.). (1989). *Rational analysis for a problematic world: Problem structuring methods for complexity, uncertainty and conflict*. Chichester: Wiley.

Rosenhead, J., & Mingers, J. (Eds.). (2001). *Rational analysis for a problematic world revisited: Problem structuring methods for complexity, uncertainty and conflict.* Chichester: Wiley.

Rosenhead, J., & Thunhurst, C. (1982). A materialist analysis of operational research. *Journal of the Operational Research Society, 33*, 122–133.

Schon, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. San Francisco: Jossey-Bass.

Vidal, R. V. V. (2004). Special issue on applications of problem structuring methods. *European Journal Operational Research, 152*, 631–640.

Processor Sharing

A queueing discipline whereby the server shares its effort over all customers present.

See

▶ Queueing Theory

Product Form

▶ Product-Form Solution

Product Form of the Inverse (PFI)

The inverse of a matrix expressed as the product of sequence of matrices. The matrices in the product are elementary elimination matrices.

See

▶ Eta File
▶ Simplex Method (Algorithm)

Product-Form Solution

When the steady-state joint probability of the number of customers at each node (station) in a queueing network is the product of the individual probabilities times a multiplicative constant, as in $\Pr\{N_1 = n_1, N_2 = n_2, \ldots, N_J = n_J\} = K\pi(n_1)\pi(n_2)\ldots\pi(n_J)$, the network is said to have a product-form solution. Sometimes the designation of a product-form solution requires that the multiplicative constant K also decompose into separate factors for each node, as holds for open Jackson networks but not for closed Jackson networks. Variants of such product-form solutions also occur in some non-network queues, such as those with vacations.

See

▶ Networks of Queues
▶ Queueing Theory

Product-Mix Problem

▶ Activity-Analysis Problem
▶ Blending Problem

Production Function

▶ Economics and Operations Research

Production Management

Jaya Singhal[1], Gabriel R. Bitran[2] and Sriram Dasu[3]
[1]University of Baltimore, Baltimore, MD, USA
[2]Massachusetts Institute of Technology, Cambridge, MA, USA
[3]University of Southern California, Los Angeles, CA, USA

Introduction

Some of the important objectives of a manufacturing system are to produce in a timely manner products that conform to specifications, while minimizing costs. The strategic measures of performance of a manufacturing system are cost, quality, flexibility, and delivery. Often hundreds of products are produced by a facility, and the entire production process may span several facilities that are geographically dispersed. In many industries the production network consists of plants that are located in different countries.

Production management entails many decisions that are made at all levels of the managerial hierarchy. Manufacturing processes involve a large number of people in many different departments and organizations, and utilize a variety of resources. In addition to the quality of human resources employed, operational efficiency depends upon the location and capacity of the plants, choice of technology, organization of the production system, and planning and control systems used for coordinating the day-to-day activities. The complexity of the problems associated with effectively and efficiently utilizing all the resources — manpower, machines, materials — needed for producing goods often necessitates the development of mathematical models to aid decision making.

Manufacturing decisions can be classified into three categories: strategic, tactical and operational. Strategic decisions pertain to decisions such as degree of vertical integration, items to produced inhouse, size and location of facilities, choice of technology, nature of equipment (general versus special purpose), long-term raw material and energy contacts, skills of employees, organization design, and so forth, that have long-term consequences and can not be easily reversed. Tactical decisions have shorter horizons of 6 month to 2 years. They include decisions such as aggregate production planning (levels of production and inventory, work force, and subcontracting), facility layout, and incremental capacity expansion. Operational decisions pertaining to issues such as order processing, detailed production scheduling, follow up, maintenance routines, and inventory control rules, drive the day to day activities.

The nature of the problems faced by a production manager depends on the characteristics of the market that the facility is competing in. For this reason it is useful to distinguish between different types of manufacturing systems. The variety and volume of products produced are critical for determining the type of the manufacturing system. Manufacturing systems have been classified into job shops, batch shops, flow lines and continuous processes on the basis of the volume and variety of the product mix. Job shops produce many different products in small quantities, each with different processing requirements. Typically the products are customized and are made only after receiving an order. At the other end of the spectrum are flow lines and continuous processes that produce a limited number of products in very high volumes. Demand is met from finished goods inventories. Batch shops lie in between these two extremes. Models for aggregate production planning are described first, followed by the models for job shops, batch shops, flow lines and continuous processes. Hopp and Spearman (2000) provide a detailed coverage of these and related topics.

Aggregate Production Planning is concerned with determination of the levels of production, inventory, work force, and subcontracting to respond to fluctuating demand. With a stable work force, the level of production can be changed by using over-time or undertime. The size of work force can be varied by hiring and layoff. Fluctuating demand can also be met by accumulating seasonal inventory. An organization may also have the option of backordering or losing sales. The relevant costs are for: (1) regular payroll and overtime; (2) carrying inventory; (3) backordering or lost sales (including the possible loss of customer goodwill, lost revenue, and penalties for late delivery): and (4) hiring (including training and learning) and layoff.

Real-world production planning may involve as many as 10,000 products (Hax and Candea 1984). With 10 decision periods, this can mean more than

100,000 variables. If the number of units sold is also a decision variable, the problem may involve more than 200,000 variables. Here quadratic and linear cost models are described. Hwang and Cha (1995), Nam and Logendran (1992), Silver et al. (1997), Thomas and McClain (1993), and Venkataraman and Smith (1996) have discussed other models and methodologies. Penlesky and Srivastava (1994) described the use of spreadsheets for production planning.

Quadratic cost models — Models with quadratic costs have several major advantages. They allow for a realistic cost structure in the planning process. They also allow uncertainties to be handled directly since they minimize the expected cost if unbiased expected demand forecasts are given (Hax and Candea 1984, p. 88; Simon 1956). The resulting solution is fairly insensitive to large errors in estimating cost parameters (Hax and Candea 1984). Hax and Candea also pointed out that this is an attractive property because of the difficulty in providing accurate cost.

The production and work force smoothing model developed by Holt et al. (1960) consists of a quadratic cost function constrained by linear equations to balance production, inventory, and sales. It selects production and work force levels in each of T periods so as to satisfy demand forecast while minimizing the sum of the costs over the T periods. Let P_t, W_t, I_t, and D_{t} represent production volume, work force level, end of period inventory, and demand forecast for period t, where the initial inventory and work force are given. The cost in period t consists of the following components:

Regular payroll costs :	C_1W_t
Hiring and layoff costs :	$C_2(W_t - W_{t-1} - C_{11})^2$
Overtime costs :	$C_3(P_t - C_4W_t)^2 + C_5P_t -$ $C_6W_t + C_{12}P_tW_t$
Inventory related costs :	$C_7(I_t - C_8 - C_9D_t)^2$

The model may be formulated as:

$$\begin{aligned}\text{Minimize Z} \ = \ & \sum_{t=1}^{T}[(C_1 - C_6)W_t + C_2(W_t - W_{t-1} - C_{11})^2 \\ & + C_3(P_t - C_4W_t)^2 + C_5P_t \\ & + C_{12}P_tW_t + C_7(I_t - C_8 - C_9D_t)^2]\end{aligned} \tag{1}$$

subject to:

$$P_t - D_t = I_t - I_{t-1} \tag{2}$$

Holt et al. focused on an infinite planning horizon with stationary costs and derived the following two linear decision rules for the first period:

$$P_1 = \theta_1 + \theta_2 I_0 + \theta_3 W_0 + \sum_{t=1}^{T} \phi_t D_t$$

$$\text{and } W_1 = \theta_4 + \theta_5 I_0 + \theta_6 W_0 + \sum_{t=I}^{T} \mu_{\mathrm{t}} D_{\mathrm{t}},$$

where $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_t$, and μ_t ($t = 1, 2, \ldots, T$) are functions of the cost coefficients. The infinite series can be truncated after an appropriate number of periods T.

Singhal and Singhal (1996) developed simple computational procedures for finite horizon cases. These can be used for arbitrary time-varying cost coefficients. The complexity of the procedures grows only linearly with T. They generate the values of production, work force, and inventory levels for each period in the planning horizon. Finally, the procedures lend themselves to sensitivity analysis with respect to terminal values and to generate alternate plans.

It is beneficial to generate a collection of alternate plans on the basis of alternative terminal conditions and evaluate them more precisely according to the actual cost structure. This is usually more complex than the quadratic cost function used in the Holt et al. model. Sensitivity analysis can also be used to eliminate plans that may include negative values of P_t, W_t, or I_t. If only I_T is specified, one can compute Z as a simple quadratic function of $W_T : Z = h + kW_T + mW\ ^2_T$ where h, k, and m are functions of the cost coefficients, ending inventory, and demand forecasts. The optimum value of W_T is then easily computed as $W_T = -k/2m$. If W_T, rather than I_T, is specified, then Z can be obtained as a quadratic function of I_T. If the terminal condition is not specified for any variable, one can obtain Z as a quadratic function of both W_T and I_T (or P_T).

One can compute optimal plans for a menu of combinations of terminal values (I_T, W_T) so as to create a menu of alternative plans which can be evaluated in more detail with respect to alternative cost structures, constraints, and objectives. The alternate plans provide considerable flexibility to the decision maker because they can be evaluated in the context of (a) constraints not included in the model, (b) actual costs, and (c) implications beyond the planning horizon.

Constraints not included in the model — The model does not specify that P_t, W_t, or I_t be non-negative. The solution approaches developed by Holt et al. (1960) or Singhal and Singhal (1996) do not guarantee it either. However, the values of θ_1, θ_2, θ_3, θ_4, θ_5, θ_6, θ_t, and μ_t ($t = 1, 2, \ldots, T$) in the decision rules for an actual problem (Holt et al. 1960) indicate that for most problems, they will be nonnegative. For cases where a solution may include negative values of P_t, W_t, or I_t, sensitivity analysis can be used to determine the ranges of the terminal boundary conditions for which all values of P_t, W_t, and I_t are non-negative. If implementation of the optimal solution is difficult because of extremely low or extremely high levels of inventory, production, or work force in some periods, trade-offs can be made between the additional cost and the ease of implementation of alternate plans that are within the constraints on inventory, production, and work force.

Actual costs — The costs of various plans, including the optimal plan refer to the costs approximated by the linear quadratic model, not to the actual costs. In testing the model for a real-world problem, one may obtain actual costs for one or more alternate plans that are lower than those of the optimal plan.

Implications beyond the planning horizon — The organization may anticipate or plan some changes beyond the planning horizon of the model. For example, it may retire workers or introduce technology that requires fewer or more workers. If the organization plans to introduce technology that requires fewer workers, it would choose a plan that would require a smaller work force towards the end of the planning horizon. Similarly, if some workers are expected to retire in the near future, the organization would choose a plan that would require hiring more workers towards the end of the planning horizon. The exact choice of the plan will depend on the magnitude of the changes beyond the planning horizon and the cost penalty during the planning horizon. In some cases, the optimal levels of inventory and work force in the final period may be incompatible with the demand forecasts for periods beyond the planning horizon (these forecasts may be too imprecise to extend the length of the planning horizon but they may indicate the overall magnitude of demand). In such cases, trade-offs can be made by comparing the possible benefits of an alternate plan and the cost penalty associated with it.

Both the finite and infinite horizon versions can be implemented on the rolling basis. In the infinite horizon version, no consideration is given to information beyond a certain period. In the finite horizon version, the implications beyond the planning horizon are first included in the specification of the terminal conditions and then evaluated through sensitivity analysis.

Bergstrom and Smith (1970) extended the Holt et al. model to a multi-product situation. It is given as

$$\begin{aligned}\text{Minimize TC} = \\ \sum_{t=1}^{T}\Bigg[&C_1W_t + \sum_{i=1}^{N}\left[C_{i7}(I_{it} - C_{i8} - C_{i9}D_{it})^2\right] \\ &+ C_3\left(\sum_{i=1}^{N}k_iP_{it} - C_4W_t\right)^2 \\ &+ \sum_{n=1}^{N}C_5k_iP_{it} - C_6W_t \\ &+ C_{12}W_t\left(\sum_{i=1}^{N}k_iP_{it}\right) \\ &+ C_2(W_t - W_{t-1} - C_{11})^2\Bigg]\end{aligned}$$

subject to:

$$I_{it} = I_{i,t-1} + P_{it} - D_{it}, \quad i = 1, 2, \ldots, N; \; t = 1, 2, \ldots, T,$$

where N and T denote the number of products and periods respectively; P_{it}, D_{it}, and I_{it} represent the production, demand forecast, and inventory of product i during period n; k_i represents the standard labor time to complete one unit of product i; and W_t represents the work force during period t. The C_i are the cost coefficients. Aggregate production L_t, aggregate inventory I_t, and aggregate demand forecast D_t can be written as

$$L_t = \sum_{i=1}^{k_i} P_{it}, \quad t = 1, 2, \ldots, T$$

$$I_t = \sum_{i=1}^{k_i} I_{it}, \quad t = 1, 2, \ldots, T$$

$$D_t = \sum_{i=1}^{k_i} D_{it}, \quad t = 1, 2, \ldots, T$$

$$I_{it} = I_{t-1} + L_t - D_t, \quad t = 1, 2, \ldots, T$$

All decision variables are unconstrained. Initial conditions I_0, W_0, and I_{0i} (i = 1, 2,…, N) and the final conditions (work force and aggregate inventory) are specified. Singhal (1992) developed a simple and efficient non-iterative algorithm for obtaining the optimal values of the levels of production management in, inventory, and work force during the planning horizon. The efficiency is achieved by exploiting the special structure of the recurrence relations obtained by differentiating the cost function. Once the input data are developed, the computation time will remain the same irrespective of the number of products which, as noted earlier, could be as many as 100,000.

Linear cost models — Linear programming models are widely used because they can be easily tailored to a specific situation. Many constraints can be directly included in the model. A major advantage of linear programming models is the availability of computer codes that can solve very large problems. Most cost structures are generally linear within the range of interest. If they are not, one can use linear approximations. Another advantage is parametric and sensitivity analyses. The dual solution can be used to obtain the costs of constraints and one can easily perform sensitivity analysis on cost parameters and demand forecasts. For a more detailed discussion of linear programming models, see Hax and Candea (1984) and Silver et al. (1997). Hax and Candea (1984) described the following general purpose model:

$$\text{Minimize } Z = \sum_{i=1}^{N} \sum_{t=1}^{T} \left(d_{it} P_{it} + c_{it} I_{it}^{+} + b_{it} I_{it}^{-} \right) + \sum_{t=1}^{T} \left(w_t W_t + o_t O_t + h_t H_t + f_t F_t \right)$$

subject to:

$$P_{it} + I_{i,t-1}^{+} - I_{i,t-1}^{-} - D_{it} = I_{it}^{+} - I_{it}^{-} \qquad i = 1, 2, \ldots, N;$$
$$t = 1, 2, \ldots, T,$$
$$W_t - W_{t-1} = H_t - F_t \qquad t = 1, 2, \ldots, T,$$
$$O_t \le p W_t \qquad t = 1, 2, \ldots, T,$$
$$P_{it}, I_{it}^{+}, I_{it}^{-}, W_t, O_t, H_t, F_t \ge 0$$
$$i = 1, 2, \ldots, N; \qquad t = 1, 2, \ldots, T$$

P_{it} = Units of item i to be produced in period t
D_{it} = Forecast demand for item i in period t
d_{it} = Cost of producing one unit of product i in period t
c_{it} = Cost of carrying one unit of inventory of product i from period t to t + 1
b_{it} = Cost of backordering one unit of inventory of product i from period t to t + 1
w_t = Cost of one regular labor hour in period t
W_t = Regular labor hours employed in period t
o_t = Cost of one overtime labor hour in period t
O_t = Overtime labor hours used in period t
h_t = Cost of hiring one labor hour in period t
H_t = Labor hours of regular work force hired in period t
f_t = Cost of laying off one labor hour in period t
F_t = Labor hours of regular work force laid off in period t
i_{it}^{+} = Inventory of product i at the end of period t
I_{it}^{-} = Units of product i backordered at the end of period t
p = An upper bound on overtime as a fraction of regular hours

The first constraint is similar to the production-inventory balance equation in the linear-quadratic model when $I_{it} = I_{it}^{+} - I_{it}^{-}$, t = 1, 2,…, T. The second constraint shows the changes in the level for work force due to hiring and layoff. The third constraint provides a limit on the overtime; the limit is proportional to the level of work force.

Job Shops

Job shops specialize in producing customized products, and the production process has the flexibility to produce many different products. Due to the high variety the flows in job shops are jumbled, thus making it very difficult to predict and manage the completion times of jobs. Since most of the jobs are produced after receiving an order from a customer, very important managerial tasks are to accurately predict due dates, ensure that the quoted dates are not violated, and use resources effectively and efficiently.

Operational Problems — The challenge of managing day to day operations has given rise to a rich set of combinatorial optimization problems. The most basic operational problem is to determine a schedule that specifies when each job will be allocated different resources. Associated with each job are the arrival time, a due date and a set of operations. Each operation requires a set of resources for some duration, and there may be precedence constraints on the order in which the operations can be performed.

A variety of performance measures have been considered for evaluating alternative schedules. Common performance measures are the average or maximum time a set of jobs remains in the facility, number of jobs that are late, or the average or maximum tardiness for a set of jobs. Most of the problems of job shop schedule optimization problems, except for a small class, are computationally intractable (Lenstra et al. 1977; French 1982). Hence for most practical problems the emphasis has been on heuristics.

Researchers have successfully analyzed job shops with special structures. Many insights have been gained into the single machine and single stage, multiple machine scheduling problems. For multiple stage job shops, analysis has been possible, provided all the jobs follow the same route.

Job shop scheduling models can be classified into static and dynamic models. In static models the set of requirements including job arrival times and processing requirement are known in advance. In contrast, in dynamic job shop models new arrivals are permitted. The arrival times may be stochastic and the processing requirements may also vary dynamically.

Mathematical programming approaches have been employed to study static job shop problems. For performance measure that are non-decreasing in the completion time of the job, dynamic programming techniques have been employed to generate optimal solutions for problems of modest size. Dynamic programming based approaches have also been useful in identifying dominance criteria to reduce the number of schedules to be evaluated. Several heuristics have been developed that exploit dominance criteria. Integer programming formulations of scheduling problems have also been used to generate near optimal solutions. Typically some complicating constraints in the integer program are relaxed to yield tractable sub-problems.

While most of the theory focuses on static job shop models that assume deterministic requirements, most practical problems are dynamic and stochastic. For such complex environments analysis has largely been restricted to simulations of local dispatching rules. Each station employs a dispatching rule — for example, process jobs in increasing order of processing times — and the overall performance of the shop is evaluated via Monte Carlo simulations. Many dispatching rules have been discussed in the literature. Further details regarding scheduling algorithms are given in Conway et al. (1967), Graves (1981), and O'Eigeartaigh et al. (1985).

An important development in the area of scheduling dynamic shops has been to approximate the job shop scheduling problem by a Brownian control problem. Although the size of the networks analyzed is small, since the focus is on bottleneck stations the method is useful in many practical situations. The Brownian control problems have been useful in identifying near optimal scheduling policies for minimizing the average lead times (Wein 1990).

Strategic and tactical problems — Since most of the operational problems of sequencing and scheduling jobs through a shop floor are computationally intractable, there is a need to design the job shops such that simple real time control rules are adequate to obtain good performance. The long term performance of the shop will depend on the types of jobs processed by the facility (product mix), the capacity and technology of different stations, and the rules employed to quote due dates and manage the flow through the shop floor. Tactical and strategic decisions regarding each of these variables require models that predict the medium to long term performance of the job shops.

One approach for assessing the long term performance is to employ Monte Carlo simulations. The strength of simulation models lies in their ability to incorporate many features, such as (i) complex control rules — for example, local dispatching rules, control of input to the shop, etc.; (ii) complex arrival patterns — for example, correlated demands, non-stationary demand, etc.; and (iii) complex resource requirements and availability — for example, multiple resources, machine failures, etc. A broad range of performance measures can also be assessed through simulation models. These models, however, are time consuming and cannot identify optimal parameters for the policies being investigated.

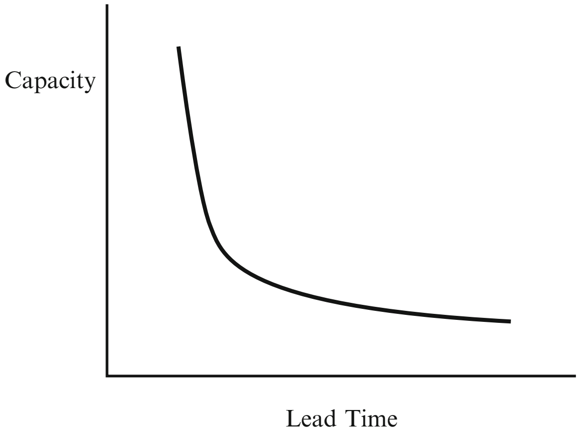
Open queueing network models have been proposed to evaluate the long term performance of job shops. Good approximation procedures have been developed to estimate the average queue lengths in networks with features such as general processing and interarrival time distributions, multiple job classes, and class dependent deterministic routing through the network.

An approximation procedure that has been frequently employed is the parametric decomposition approach (PDA). Under the PDA, each node is treated as being stochastically independent and all the performance measures are estimated based on the first two moments of the inter-arrival and service time distributions at each node. Extensive testing has shown that PDA provides accurate estimates of the average queue length at each node in very general networks. Limitations of the approach are that all the measures are for steady state, only the average queue lengths are accurately predicted, and the analysis is based on the assumption that the jobs are processed on a first come first served basis. Nevertheless, the power of this approach lies in the ease with which complex networks can be analyzed, which in turn facilitates the design of networks.

The PDA has enabled the analysis of several optimal facility design problems. One such problem is:

- Objective: Minimize total cost of equipment.
- Decision Variables: Capacity of each station in the network, and technology.
- Constraints: Upper bounds on the average lead time for different job classes.

This model addresses the relationship between average lead times and the choice of equipment. Since system design is based on multiple criteria, it is useful to develop curves that reflect the trade-off between lead times and cost of equipment. This can be done by parametrically varying the upper bound on the permissible lead times. Figure 1 provides a possible trade-off curve (Bitran and Tirupati 1989). Details regarding the application of queueing models to job shops are given in Bitran and Dasu (1992).



Production Management, Fig. 1 Illustrative trade-off curve

Batch Shops

The variety of jobs processed in a batch shop is less than that in job shops; furthermore, the set of products that are produced by the facility may be fixed. Nevertheless, the production volume of each product is such that several products may share the same equipment. Often the demand for final goods is met from finished goods inventory and production plans are based on demand forecasts. A large number of discrete part manufacturing systems can be classified as batch shops.

Operational problems: The time and cost for switching machines from one product to the next poses one of the biggest problems in managing batch shops. Although job shops can also have significant set-ups, since each job is unique the set-up time can be incorporated in that job's total processing time. On the other hand, in batch shops, the same products are produced repeatedly and there is an opportunity to mitigate the effect of set-ups by combining or splitting orders. Consequently much attention has been paid to problems of determining batch quantity of and the sequence in which each item is produced. The primary trade-offs are between inventory carrying, shortage and set-up costs.

A classic lot sizing problem is the economic lot scheduling problem (ELSP). The ELSP seeks the optimal lot size at a single production stage when the demand rate for each item is fixed and deterministic (Panwalker and Iskander 1977). The objective of the analysis is to determine the frequency with which each item is to be produced so as to minimize the average set-up and holding costs without ever stocking out.

Many of the solution procedures for ELSP consist of three steps. First, ignoring the capacity constraint, the optimal production frequency for each item is determined. Next the frequencies are rounded off to an integer multiple of a base period. In the final step a solution that specifies the sequence in which each item is produced is generated. Roundy (1986) showed that in the second step if the integer multiple is restricted to some power of 2, then a near optimal solution can be found. In recent years researchers have begun to extend the approaches developed for ELSP to multistage multi-machine problems.

ELSP is a continuous time model. In practice production plans are made on a periodic basis, prompting several researchers to develop and analyze discrete time models of the lot-sizing problems. Below a single-stage, multi-item, multi-period, capacitated lot-sizing problem is formulated:

$$\text{Minimize } \sum_{t=1}^{T}\left\{p_t(X_t)+h_t(I_t)+\sum_{i=1}^{I}s_{it}\delta(X_{it})\right\}$$

subject to:

$$I_{i,t-1}+X_{it}-I_{it}=D_{it}\quad t=1,2,\ldots,T;\ \ i=1,2,\ldots,I.$$

$$\sum_{i=1}^{I}X_{it}\le X_t\quad t=1,2,\ldots,T.$$

$$\delta(X_{it})=\begin{cases}1 & \text{if } X_{it}=0\\ 0 & \text{otherwise}\end{cases}$$

$$I_{it},X_{it}\ge 0\quad t=1,2,\ldots,T;\ \ i=1,2,\ldots,I.$$

where X_{it}, I_{it}, C_t, D_{it} and s_{it} denote respectively for period t and product i, the production quantity, the ending inventory, the capacity, the demand, and the setup cost; X_{it} and I_{it} are the only decision variables; and X_t and I_t are vector with elements $\{X_{it}\}$ and $\{I_{it}\}$, respectively. The functions $p_t(\cdot)$ and $h_t(\cdot)$ denote respectively the variable production and inventory holding costs.

Once again, except for a small class, the lot-sizing problems are NP-hard (Garey and Johnson 1979); Bitran and Yanasse 1982). The following two lot-sizing problems, however, can be solved in polynomial time and have been the basis of many approximation procedures: (a) the single item lot-sizing problem without capacity constraints, and concave variable production and inventory holding costs; and (b) single item problem, with constant capacity, and concave variable production and inventory holding costs.

Multistage systems producing multiple products with dynamic demands, usually require extensive information and considerable computational effort to find optimal solutions. For these reasons, hierarchical planning systems have been proposed. At the highest level in the hierarchy an aggregate plan with a horizon of several, usually 12, months is developed. If the demand is seasonal, the horizon should cover the full demand cycle. Over such horizons it is impractical to obtain detailed information about demand for each item and the availability of every resource. Hence, it becomes necessary to aggregate the items into families, and the machines into machine centers, etc. The aggregate plan determines the time phased allocation of aggregate resources to different part families. The plan focuses on the primary trade-offs among the cost of varying production resources employed by the firm, the costs of carrying inventory (and possibly backordering demand), and major setup costs. The extended horizon enables the facility to respond to seasonality in demand.

The aggregate plan becomes the basis for determining the detailed production schedule for each item. The detailed resource allocation decisions are constrained by the decisions made at the aggregate planning level.

The number of hierarchical planning stages, the degree of aggregation at each level, and the planning horizon lengths affect the quality of the plan and must be carefully determined for each context. Many researchers have studied hierarchical planning systems. Bitran and Tirupati (1993) and Hax and Candea (1984) contain discussions of this approach.

Once the plans have been disaggregated and the monthly requirements of each item are known, there are a number of approaches for scheduling and controlling the flow of the items through the shop. One approach is to time the release of the orders to the shop so that the required quantities of the items become available by the date specified by the hierarchical planning system. In this approach, also referred to as the push system, an estimate is made of production lead times, and order releases are offset by the lead times. The scheduling decisions at each work station may be made on the basis of the queue in front of each work station. Scheduling models developed for job shops are also useful here.

An alternate approach for operating the shop is the pull system. Under this approach the work-in-process inventory level after a production stage determines the production decisions at that stage. The buffer inventories are maintained at planned levels and a production order is triggered if the inventory level drops below the threshold.

Since the push system operates on the basis of planned lead times, OR/MS models have been developed to understand the relationship between release rules, capacity and lead times. The key decision variable in pull systems is the size of each buffer. Several researchers have examined the impact of buffer sizes on the shop performance (Conway et al. 1988).

Strategic and Tactical Problem — An approach advocated for simplifying the operations of batch shops is to partition the facility into cells. Parts produced by the facility are grouped into families and each family is assigned to a cell. Ideally all operations required for a family of parts are performed in the same cell. The advantages of cellular manufacturing systems are simplified flows, and reduced lead times and setup costs. These benefits may be partially offset by the need for additional equipment. Many different criteria — such as part geometry, production volumes, setups, and route through the shop — have been proposed for forming part families. Researchers have also investigated several algorithms for identifying alternative partitions. Typically these algorithms begin with a product-process matrix. In this matrix rows correspond to parts and columns correspond to machines. An element *ij* in this matrix is one if a part *i* requires a machine *j* and zero otherwise. The columns and rows of the matrix are interchanged so as to produce a block diagonal matrix. Each block identifies a set of resources and jobs that does not interact with the remaining operations, and so corresponds to a cell.

As in the case of job shops, batch shops system design can be improved if the medium to long term performance of the shop can be assessed. Closed and open queueing network models and simulation based models are useful for assessing the long term performance of batch shops. The objective of these models is to determine the relationship among capacity of different cells, lot sizes, and lead times (Bitran and Dasu 1992).

Queueing network models assume that the processing rate at each station is fixed. In practice the processing rate at each station may vary. Variations may be due either to the allocation of additional (human) resources to a stage or simply because the queue length has a motivational effect on the machine operator. Based on these observations, in recent years an alternative class of tactical models of the shop have been proposed (Graves 1986). Here the production rates are assumed to vary as a function of the size of the queue length. The processing rates at each stage are allowed to vary so as to ensure that the time spent at a station is the same for every job. The model therefore enables managers to plan the lead times for each stage.

Flow Lines and Continuous Operations

Included in this class are all systems that are dedicated to the production of (one or few) items in large volumes. Examples of such systems include assembly lines, transfer lines, and continuous operations such as cement and oil derivatives manufacture. The demand is often met from finished goods inventory and thus the main focus tends to be on the management of the corresponding inventory levels and the supply chain. The operational problems are relatively simple and are omitted.

Tactical problems — An important operational problem is to manage the trade-off between the cost of varying the production rate and the cost of finished goods inventory. The aggregate planning models discussed earlier are applicable here. Typically, all the stages of the production system have equal capacity, hence, managing the flow through the facility does not pose a significant problem. In assembly lines, the balance is achieved by carefully assigning tasks to different work stations — a complex combinatorial optimization problem. Several algorithms have been developed for assembly line balancing.

Strategic problems — High volume production systems frequently compete on the basis of low costs and supply large geographically dispersed markets. It is therefore not uncommon to have many plants that cater to different markets. OR/MS models have been developed to aid in the design of the multiplant networks and the distribution systems (Erlenkotter 1978; Federgruen and Zipkin 1984). Here the discussion is restricted to the plant location problems.

The number of plants their capacity and location have a big effect on production and distribution costs. Models have been developed to analyze the trade-off between the fixed costs of setting up plants and the variable (transportation and production) costs of operating the plants. The models assume that a set of markets with known demands have to be supplied and the decision variables are the number of plants, their location and capacities (Erlenkotter 1978).

Concluding Remarks

Production management involves many complex trade-offs. As a result many mathematical models have been developed to aid decision makers. This is certainly not an exhaustive list and excludes many important problem areas such as inventory management, preventive maintenance, capacity expansion, and quality control. The focus has been on models that are concerned with the flow of goods through a manufacturing system. Even within this domain, in order to provide a broad overview, many important models that deal with specialized systems were not discussed, such as intelligent manufacturing systems.

The problems arising in each type of production system were described as if each plant operated in isolation. In practice, a production system is likely to consist of a network of plants. While some plants may be batch or job shops others are likely to be assembly or continuous processes. The problems of coordinating these networks was not discussed.

Most of the OR/MS models focus on managing the trade-offs among setup costs, inventory carrying costs and cost of varying production rates. On the other hand, many gains in productivity are due to the elimination (or mitigation of) the factors that give rise to these trade-offs. For example reduction in set-up costs and times reduces lead times, increases the ability of the system to produce a wider mix of products, diminishes the role of inventories and simplifies the management of batch shops. Researchers have begun to develop models that quantify the benefits of and guide such process improvement efforts (Porteus 1985; Silver 1993).

See

- ▶ Automation in Manufacturing and Services
- ▶ Dynamic Programming
- ▶ Facilities Layout
- ▶ Facility Location
- ▶ Flexible Manufacturing Systems
- ▶ Hierarchical Production Planning
- ▶ Integer and Combinatorial Optimization
- ▶ Inventory Modeling
- ▶ Job Shop Scheduling
- ▶ Location Analysis
- ▶ Operations Management
- ▶ Queueing Theory
- ▶ Supply Chain Management

References

Bergstrom, G. L., & Smith, B. E. (1970). Multi-item production planning — An extension of the HMMS rules. *Management Science, 16*, 614–629.

Bitran, G. R., & Dasu, S. (1992). A review of open queueing network models of manufacturing systems. *Queueing Systems: Theory and Applications, 12*, 95–134.

Bitran, G. R., & Tirupati, D. (1989). Trade-off curves, targeting and balancing in manufacturing networks. *Operations Research, 37*, 547–564.

Bitran, G. R., & Tirupati, D. (1993). Hierarchical production planning. In S. C. Graves, A. H. G. Rinnooy Kan, & P. Zipkin (Eds.), *Logistics of production and inventory* (Handbooks in O.R. and M.S, Vol. 4). Amsterdam: Elsevier Science Publishers.

Bitran, G. R., & Yanasse, H. H. (1982). Computational complexity of capacitated lot sizing problem. *Management Science, 28*, 1174–1186.

Burbridge, J. L. (1979). *Group technology in the engineering industry*. London: Mechanical Engineering Publications.

Conway, R. W., Maxwell, W., McClain, J. O., & Thomas, L. J. (1988). The role of work-in-process inventory in serial production lines. *Operations Research, 36*, 229–241.

Conway, R. W., Maxwell, W. L., & Miller, L. W. (1967). *Theory of scheduling*. Reading, MA: Addison-Wesley.

Erlenkotter, D. (1978). A dual-based procedure for uncapacitated facility location. *Operations Research, 26*, 992–1005.

Federgruen, A., & Zipkin, P. (1984). Approximation of dynamic multi-location production and inventory problems. *Management Science, 30*, 69–84.

French, S. (1982). *Sequencing and scheduling: An introduction to the mathematics of the job-shop*. New York: John Wiley.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of N.P. completeness*. San Francisco: Freeman.

Graves, S. C. (1981). A review of production scheduling. *Operations Research, 29*, 646–675.

Graves, S. C. (1986). A tactical planning model for job shop. *Operations Research, 34*, 522–533.

Hax, A. C., & Candea, D. (1984). *Production and inventory management*. New Jersey: Prentice-Hall.

Holt, C. C., Modigliani, F., Muth, J. F., & Simon, H. A. (1960). *Planning production, inventories, and work force*. Englewood Cliffs, NJ: Prentice-Hall.

Hopp, W. J., & Spearman, M. L. (2000). *Factory physics* (2nd ed.). New York: Irwin/McGraw Hill.

Hwang, H., & Cha, C. N. (1995). An improved version of the production switching heuristic for the aggregate production planning problem. *International Journal of Production Research, 33*, 2567–2577.

Lenstra, J. K., Rinnooy Kan, A. H. G., & Brucker, P. (1977). Complexity of machine scheduling problems. *Annals of Discrete Mathematics, 1*, 343–362.

Nam, S. J., & Logendran, R. (1992). Aggregate production planning — A durvey of models and methodologies. *European Journal of Operational Research, 61*, 255–272.

O'Eigeartaigh, M., Lenstra, J. K., & Rinnooy, A. H. G. K. (1985). *Combinatorial optimization — Annotated bibliographies*. New York: John Wiley.

Panwalker, S. S., & Iskander, W. (1977). A survey of scheduling rules. *Operations Research, 25*, 45–61.

Penlesky, R., & Srivastava, R. (1994). Aggregate production planning using spreadsheet software. *Production Planning & Control: The Management of Operations, 5*, 524–532.

Porteus, E. L. (1985). Investing in reduced setups in the EOQ model. *Management Science, 31*(8), 998–1010.

Roundy, R. (1986). A 98% effective lot-sizing rule for a multi-product, multi-stage production/inventory system. *Mathematics of Operations Research, 11*, 699–727.

Silver, E. A. (1993). Modeling in support of continuous improvements towards achieving world class operations. In R. Sarin (Ed.), *Perspectives in operations management: essays in honor of Elwood S. Buffa*. Norwell, MA: Kluwer.

Silver, S. A., Pyke, D. F., & Peterson, R. (1997). *Inventory management and production planning and scheduling* (3rd ed.). New York: John Wiley.

Simon, H. A. (1956). Dynamic programming under uncertainty with a quadratic cost function. *Econometrica, 24*(1), 74–81.

Singhal, K. (1992). A noniterative algorithm for the multiproduct production planning and work force planning problem. *Operations Research, 40*, 620–625.

Singhal, J., & Singhal, K. (1996). Alternate approaches to solving the Holt et al. model to performing sensitivity analysis. *European Journal of Operational Res., 91*, 89–98.

Thomas, J., & McClain, J. O. (1993). An overview of production planning. In S. C. Graves, A. H. G. Rinnoy Kan, & P. Zipkin (Eds.), *Logistics of production and inventory* (Handbooks in O.R. and M.S, Vol. 4). Amsterdam: Elsevier Science Publishers.

Venkataraman, R., & Smith, S. B. (1996). Disaggregation to a rolling horizon master production schedule with minimum batch-size production restrictions. *International Journal of Production Research, 34*, 1517–1537.

Wein, L. M. (1990). Optimal control of a two-station Brownian network. *Mathematics of Operations Research, 15*, 215–242.

Production Rule

A mapping from a state space to an action space, generally used in modular knowledge representation. With roots in syntax-directed parsing of language, production rules comprise a basic reasoning mechanism, particularly in heuristic search.

See

- ▶ Artificial Intelligence
- ▶ Expert Systems

Program Evaluation

Edward H. Kaplan and Todd Strauss
Yale University, New Haven, CT, USA

Introduction

Program evaluation is not about mathematical programming, but about assessing the performance of social programs and policies. Does capital punishment deter homicide? Which job training programs are worthy of government support? How can emergency medical services be delivered more effectively? What are the social benefits of energy conservation programs? These are the types of questions considered in program evaluation.

Notable evaluations include the Westinghouse evaluation of the Head Start early childhood program (Cicarelli et al. 1969), the Housing Allowance experiment (Struyk and Bendick 1981), the Kansas City preventive patrol experiment (Kelling et al. 1974), and evaluation of the New Haven needle exchange program for preventing HIV transmission among injecting drug users (Kaplan and O'Keefe 1993). As these examples suggest, questions and issues deserving serious evaluation often are in the forefront of social policy debates in areas such as public housing, health services, education, welfare, and criminal justice.

Closely related to program evaluation are the activities of cost-benefit and cost-effectiveness analysis. These resource allocation methods help decision makers decide which social programs are worth sponsoring, and how much money should be invested in competing interventions. Program evaluation may be construed as an attempt to understand and estimate the benefits associated with the social program under study. While some evaluations attempt to relate these benefits to the costs of program activities, most program evaluations are viewed as attempts to measure benefits alone.

Program evaluation is often conducted by social scientists at the behest of organizations with some interest in the program, either as participants, administrators, legislators, managers, program funders, or program advocates. In such a charged atmosphere, how can OR/MS be useful? Program evaluation contributes to policy making chiefly by informing policy debate. Evaluation can be construed as an activity that produces important information for decision makers in the policy process (Larson and Kaplan 1981). Evaluation is also useful for framing issues, and for identifying and choosing among policy options. Evaluation is crucial to program administrators concerned with improving service delivery. These tasks are about gathering, analyzing, and using information. It is the orientation toward decision making that renders OR/MS particularly useful in the evaluation of public programs.

Program Components and the Scope of Evaluation

In the language of systems analysis, the components of social programs can be classified as inputs, processes, and outputs (Rossi and Freeman 1993). Inputs are resources devoted to the program, while outputs are products of the program. In this framework, program evaluation is usually about assessing a program's effects on outputs. Such evaluation is often called outcome or impact evaluation. Typically, the result of outcome evaluation is the answer to the question: Did the program achieve its goals?

In contrast to outcome evaluation, process evaluation is often referred to, perhaps pejoratively, as program monitoring. As the myriad details of real programs are classified simply as processes in monitoring studies, programs become black boxes. Such a framework is anti-operational. On the other hand, an OR/MS approach to process evaluation focuses on program operations, often with the assistance of appropriate mathematical models. Typical program evaluations too often lead to simplistic conclusions regarding which programs work. Focusing on program operations often results in understanding why some programs are successful and other programs fail. As an example, consider Larson's analysis of the Kansas City Preventive Patrol Experiment (Larson 1975). This experiment attempted to discern the impact of routine preventive patrol on important outcomes such as crime rates and citizen satisfaction, in addition to important intermediate outcomes such as response time and patrol visibility. The empirical results of this experiment resulted in several findings of "no difference" between patrol areas with supposedly low, regular, and high intensities of police preventive patrol. In contrast, Larson's application of back-of-the-envelope probabilistic models to this experiment showed that one should have expected such results due to the nature of the experimental design. He showed, for example, that one should not have expected large differences in police response times given the peculiarities of patrol assignments and call-for-service workloads evident in the experiment. The same models suggested that different experimental conditions, better reflecting police operations in other large American cities, could lead to different results.

An advantage of an OR/MS approach to program evaluation is that goals and objectives are stated as explicitly as possible. What is the purpose of the program under study, and how does one characterize good versus poor program performance? While the importance of such questions may be self-evident to OR/MS practitioners, most actors on the policy stage are not accustomed to such explicitness. The act of asking such questions is often, by itself, a contribution to policy debate. A defining feature of the OR/MS approach to problem solving is the association of one or more performance measures with program objectives. A performance measure quantifies how well a system functions. Performance measures should be measurable (computable if not actually observable), understandable, valid and reliable, and responsive to changes in program operations. Operational modeling of public programs can even yield performance measures not apparent a priori. For example, the evaluation of the New Haven needle exchange program involved a mathematical model of HIV transmission among drug injectors as modified by the operations of needle exchange (Kaplan and O'Keefe 1993). The model revealed needle circulation time, that is, the amount of time a needle is available for use by drug injectors, as a critical performance measure. Reducing needle circulation time reduces opportunities for needle sharing on a per needle basis. This reduces both the chance that a needle becomes infected, and the chance that an injection with a used needle transmits infection. Needle exchange adjusts the distribution of needle circulation times. The model uncovered a direct link between the exchange of needles and the probability of HIV transmission.

Methodologies

Much of program evaluation is qualitative in nature. Social science methods relying on field observation, case histories, and the like are often used. However, such qualitative data often fail to satisfy critics of particular social programs. In addition, qualitative data generally allow only coarse judgments about program effectiveness. While no panacea, quantitative assessment methods have become standard in evaluating social programs and policies. Assessments of program effects are often made by statistically comparing a group participating in the program to a control group. The randomized experiment is the archetype for this kind of comparison. Since true randomized experiments may be difficult to execute under real program settings, quasi-experimental designs are often used instead. Rather than randomly assigning participants to program and control groups, quasi-experimental methods attempt to find natural or statistical controls. Multiple regression, analysis of variance, or other statistical techniques are often used; Cook and Campbell (1979) is a classic reference on quasi-experimental methods.

The model-based techniques of OR/MS are also applicable to program evaluation. Decision analysis is obviously useful in prospectively selecting among policy options. Queueing theory may be used to analyze the delivery of a wide range of programs, including public housing assignments, 911 hotlines, and dial-a-ride van services for the elderly and disabled. Applied probability models are generally useful, while statistical methods are widely valued. Techniques for multicriteria optimization, data envelopment analysis, and the analytical hierarchy process may be useful in identifying tradeoffs among multiple objectives.

While it seems that a solid understanding of OR/MS modeling is useful in conducting program evaluation, OR/MS has been underutilized. For example, basic optimization techniques such as linear programming have not been widely applied, perhaps because formulating a consensus objective function is usually very difficult. Training in OR/MS is less common than training in statistics and other social sciences. Few of those who have been trained in OR/MS have chosen to concentrate their efforts in the evaluation of public programs. Thus, social program evaluation remains an important and fertile area for further development and application of OR/MS methods.

Professional Opportunities and Organizations

Departments and agencies of federal, state, and municipal government and international organizations typically have offices that perform evaluation activities. Examples include the U.S. Environmental Protection Agency's Office of Policy Planning and Evaluation, the New York City Public School's Office of Research, Evaluation, and Assessment, and the World Bank's Operations Evaluations Unit. A few large private or non-profit organizations under-take many program evaluations. Among such organizations are The Urban Institute, Abt Associates, RAND Corporation, Mathematica Policy Research, and Westat. Much program evaluation is done by academics, largely social scientists. There are opportunities for OR/MS practitioners to get involved. One outlet is the INFORMS College on Public Programs and Processes. The American Evaluation Association is an interdisciplinary group of several thousand practitioners and academics. The journal *Evaluation Review* publishes examples of quality evaluations.

See

- ▶ Cost Analysis
- ▶ Cost-Effectiveness Analysis
- ▶ Emergency Services
- ▶ Practice of Operations Research and Management Science
- ▶ Problem Structuring Methods
- ▶ Public Policy Analysis
- ▶ RAND Corporation
- ▶ Systems Analysis
- ▶ Urban Services

References

Cicarelli, V. G., et al. (1969). *The impact of head start*. Athens, OH: Westinghouse Learning Corporation and Ohio University.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Kaplan, E. H., & O'Keefe, E. (1993). Let the needles do the talking! Evaluating the New Haven needle exchange. *Interfaces, 23*, 7–26.

Kelling, G. L., et al. (1974). *The Kansas city preventive patrol experiment: Summary report*. Washington, DC: The Police Foundation.

Larson, R. C. (1975). What happened to patrol operations in Kansas City? A review of the Kansas City Preventive Patrol Experiment. *Journal of Criminal Justice, 3*, 267–297.

Larson, R. C., & Kaplan, E. H. (1981). Decision-oriented approaches to program evaluation. *New Directions for Program Evaluation, 10*, 49–68.

Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed.). New-bury Park, CA: Sage Publications.

Struyk, R. J., & Bendick, M., Jr. (1981). *Housing vouchers for the poor: Lessons from a national experiment*. Washington, DC: Urban Institute.

Program Evaluation and Review Technique (PERT)

A method for planning and scheduling a project which models uncertainties in activity by using optimistic, likely and pessimistic time estimates for each activity. PERT evolved when the U.S. Navy was developing a system to plan and coordinate the Polaris missile program (Malcolm et al. 1959).

See

- ▶ Critical Path Method (CPM)
- ▶ Network Planning
- ▶ Project Management
- ▶ Research and Development

References

Malcolm, D. G., Roseboom, J. H., Clark, C. E., & Fazar, W. (1959). Application of a technique for research and development program evaluation. *Operations Research, 7*, 646–669.

Project Management

Mark Westcombe and Graham K. Rand
Lancaster University, Lancaster, UK

Project management means different things to different people. Traditionally the domain of engineering, it has concerned itself with managing anything from small construction developments to large complex systems integration projects in defense, aerospace and other industries. A comprehensive survey of the development of project management since the 1940s and the issues involved in accomplishing projects is available in *The Management of Projects* (Morris 1997). In this period, OR/MS almost exclusively focused on the technical aspects of conforming to a contract using the iron triangle paradigm of management: to deliver a project to a pre-defined specification, on time, with an efficient use of resources within budget and with attention to safety. It accepted the project focus as the activities associated with the project lifecycle: defining scope; the work breakdown of the project plan; scheduling these activities; estimating costs; allocating resources and monitoring and controlling progress. OR/MS interested itself predominantly with techniques such as Program Evaluation and Review Technique (PERT) and the Critical Path Method (CPM).

Project management has since become ubiquitous within commercial and public sector organizations having been used to deliver organizational change (see Balogun et al. 2008). Businesses might now use project management discourse and techniques to manage anything from opening a new store to the acquisition, a merger with an international corporation or to complete an urban regeneration scheme. They may conceive projects, form project teams and appoint project managers to issues that previously would have been dealt with by managers responsible for day-to-day operations. A critique of this projectification of operational management is offered by Hodgson and Cicmil (2006).

This evolution of project management has led to new ways of thinking about projects (Winter and Szczepanek 2009) and the focus of the project manager is now more concerned with defining project success (Atkinson 1999), delivering long-term project outcomes and ensuring benefits that add value to an organization's operations (Cooke-Davies 2007). Similarly OR/MS is engaging more at a strategic level of projects, offering, in particular, ideas from systems thinking for the developing of processes rather than just techniques, such as for the project front-end (Winter 2009), negotiating project objectives amongst differing stakeholder perspectives and managing stakeholder relationships. OR/MS has also contributed significantly to the risk analysis of projects (Williams 1995) as risk management has come to the fore, including: mathematical modeling (Chapman and Ward 2002); qualitative modeling of the systemic nature of risk (Ackerman et al. 2007); the cost impact of disrupted learning curves (Howick and Eden 2001); and the use of system dynamics to model disruption and delay of projects in litigation (Eden et al. 2000). It has also concerned itself with project selection, Monte Carlo simulation of projects and project portfolio management.

Outside of OR/MS, topics of current concern include: project evaluation and improvement; strategic alignment; organizational learning; program management; project leadership; sustainability issues; partnering; project governance; and procurement (see Crawford et al. 2006). A special issue of the *International Journal of Project Management* is of particular interest (Winter et al. 2006), which reviews future trends in the field as well as explores key contemporary themes in depth. A comprehensive breakdown of all the tactical elements of project management can be found in the professional Bodies of Knowledge (Association of Project Management 2006; Project Management Institute 2008), as well as from the growing industry of professional courses and certification in project management, such as PRINCE2, which is widely used in UK public sector projects and offers a particular step by step approach to project management.

Professional association in project management is available through the Association of Project Management, Ibis House, Regent Park, Summerleys Road, Princes Risborough, Buckinghamshire, UK HP27 9LE, which publishes *The International Journal of Project Management*; and the Project Management Institute, which publishes the *Project Management Journal*. Note that the term project management, or project management skills, is often misleadingly appropriated as a term in personal development to cover such transferable skills as time management, prioritization, presentation skills, etc.

See

- ▶ Critical Path Method (CPM)
- ▶ Network Planning
- ▶ Practice of Operations Research and Management Science
- ▶ Program Evaluation and Review Technique (PERT)

References

Ackerman, F., Eden, C., Williams, T., & Howick, S. (2007). Systemic risk assessment: a case study. *Journal of the Operational Research Society, 58*, 39–51.

Association of Project Management. (2006). *APM body of knowledge*. High Wycombe, Buckinghamshire: Author.

Atkinson, R. (1999). Project management: Cost, time and quality, two best guesses and a phenomenon, it's time to accept other success criteria. *International Journal of Project Management, 17*, 337–342.

Balogun, J., Hailey, V. H., Johnson, J., & Scholes, K. (2008). *Exploring strategic change*. London: FT Prentice Hall.

Chapman, C., & Ward, S. (2002). *Managing project risk and uncertainty: A constructively simple approach to decision making*. London: Wiley.

Cooke-Davies, T. (2007). Managing benefit. In J. R. Turner (Ed.), *Gower handbook of project management* (pp. 245–259). Aldershot: Gower.

Crawford, L., Pollack, J., & England, D. (2006). Uncovering the trends in project management: Journal emphases over the last 10 years. *International Journal of Project Management, 24*, 175–184.

Eden, C. E., Williams, T. M., Ackermann, F. A., & Howick, S. (2000). On the nature of disruption and delay (D&D). *Journal of the Operational Research Society, 51*, 291–300.

Hodgson, D. E., & Cicmil, S. (2006). *Making projects critical*. Basingstoke: Palgrave.

Howick, S. M., & Eden, C. (2001). The impact of disruption and delay when compressing large projects: Going for incentives? *Journal of the Operational Research Society, 52*, 26–34.

Morris, P. W. C. (1997). *The management of projects*. London: Thomas Telford.

Project Management Institute. (2008). *A guide to the project management body of knowledge*. Newtown Square, PA: Author.

Williams, T. M. (1995). A classified bibliography of research relating to project risk. *European Journal of Operational Research, 85*, 18–38.

Winter, M. (2009). Using soft systems methodology to structure project definition. In T. M. Williams, K. Samset, & K. J. Sunnevåg (Eds.), *Making essential choices with scant information: Front-end decision-making in major projects* (pp. 125–144). London: Palgrave Macmillan.

Winter, M., Smith, C., Morris, P., & Cicmil, S. (2006). Directions for future research in project management: The main findings of a UK government-funded research network. *International Journal of Project Management, 24*, 638–649.

Winter, M., & Szczepanek, T. (2009). *Images of projects*. Farnham, Surrey: Gower.

Project SCOOP

Project SCOOP (Scientific Computation of Optimal Programs) was a research program of the U.S. Air Force from the late 1940s to early 1950s whose main objective was to study and solve Air Force programming and scheduling problems. It was while working on Project SCOOP problems that George B. Dantzig formulated the linear-programming model and developed the simplex method for solving such problems.

Projection Matrix

For a given matrix $\boldsymbol{A}$, its associated projection matrix is defined as $\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$. The matrix $\boldsymbol{P}$ projects any vector $\boldsymbol{b}$ onto the column space of $\boldsymbol{A}$.

See

▶ Matrices and Matrix Algebra

Proper Coloring

An assignment of colors to nodes in a graph in which adjacent nodes are colored differently.

See

▶ Graph Theory

Prospect Theory

A descriptive theory of decision making under uncertainty (human choice), which attempts to explain certain deviations of observed empirical behavior from expected utility theory.

See

▶ Choice Theory
▶ Utility Theory

References

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision making under risk. *Econometrica, 47*, 263–289.

Protocols

The elicitation of an expert's procedure by asking the expert to describe aloud how he or she is solving a problem, such as making a forecast or a decision.

See

- ▶ Artificial Intelligence
- ▶ Expert Systems
- ▶ Forecasting

Pseudoconcave Function

Given a differentiable function $f(\cdot)$ on an open convex set X, the function f is pseudoconcave if $f(\boldsymbol{y}) > f(\boldsymbol{x})$ implies that $(\boldsymbol{y} - \boldsymbol{x})^T \nabla f(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$ where $\boldsymbol{x} \neq \boldsymbol{y}$.

See

- ▶ Concave Function
- ▶ Quasi-Concave Function

Pseudoconvex Function

Given a differentiable function $f(\cdot)$ on an open convex set X, the function f is pseudoconvex if $-f$ is pseudoconcave.

See

- ▶ Convex Function
- ▶ Pseudoconcave Function
- ▶ Quasi-Convex Function

Pseudoinverse

- ▶ Matrices and Matrix Algebra

Pseudorandom Numbers

A sequence of values coming from a mathematical algorithm, which appears to be statistically drawn independently from a uniform distribution over the unit interval [0,1].

See

- ▶ Random Number Generators

Pseudo-Polynomial-Time Algorithm

An algorithm whose running time is technically not polynomial because it depends on the magnitudes of the numbers involved, rather than their logarithms.

See

- ▶ Computational Complexity

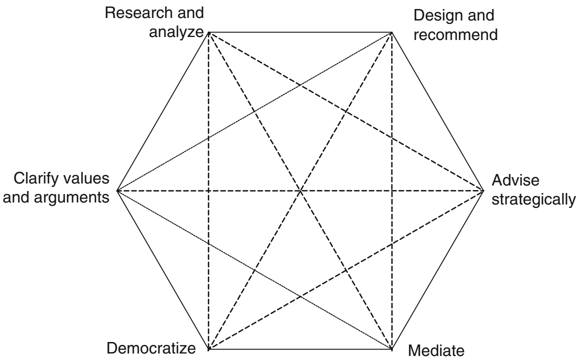
Public Policy Analysis

Warren E. Walker[1] and Gene H. Fisher[2]
[1]Delft University of Technology, Delft, The Netherlands
[2]RAND Corporation, Santa Monica, CA, USA

Introduction

Public policy analysis refers to the activities, methods, and tools that are used to give aid, advice, and support in the context of public policymaking. It covers a wide range of activities conducted with differing primary objectives and perspectives. Mayer et al. (2004) introduced a conceptual framework – the hexagon framework – that classifies the policy analysis activities in a structured manner. According to the hexagon framework, an analyst providing policy support may carry out six major clusters of activities, each having different objectives. The six objectives, represented as the vertices of the hexagon given in Fig. 1, are:

- *Research and analyze*: This type of activity aims for the generation of knowledge that can be used later for policy purposes. The major objective is to understand certain policy-relevant phenomena, and develop insights about them.
- *Design and recommend*: In certain situations the analyst can assist the decision-making process by designing alternative solutions to a problem and analyzing and possibly weighing the consequences



Public Policy Analysis, Fig. 1 Overview of objectives of policy analysis (Mayer et al. 2004)

of these alternative solutions. The main question here is more about evaluating a set of interventions, or changing the system that is related to the already known phenomena. In other words, there is a certain action orientation that ends with a policy choice or recommendation.

- *Provide strategic advice*: In certain situations, an analysis can be a strategic, client-oriented activity. The analyst can advise the client on the most effective strategy for achieving certain goals given a certain political constellation, i.e., the environment in which the client operates, the likely counter-steps of opponents, etc.
- *Mediate*: A given policy problem generally involves multiple parties that have different views and perspectives regarding the issue. Addressing the problem and coming up with an effective (i.e., accepted by all parties) policy may require the understanding of the other parties' perspectives. Hence, the task for the policy analyst may be mediating these multiple parties and promoting communication among them within a policymaking or decision-making process.
- *Democratize*: This type of policy-analytic activity aims mainly at acquiring and maintaining the involvement of all related parties in the policy process in order to make it as democratic as possible. This includes assuring the flow of proper information to all stakeholders, and the provision of opportunities for them to have their say regarding the policy issue.
- *Clarify arguments and values*: The main objective of this type of policy analysis activity is the elicitation of mindset, norms, and values of the stakeholders involved in the problem at hand. In these situations, the analyst can support or help move forward the decision-making process by analyzing the values and argumentation systems that underpin the social and political debate.

In real-life cases and projects, a policy analyst will combine one or more of these activities, albeit not all at the same time. Traditional policy analysis is focused on the 'design and recommend' vertex (see Walker 2000). The approach related to this objective is detailed below, and expanded upon in Thissen and Walker (2013). Its primary purpose is to *assist* policymakers in choosing a preferred course of action to implement in a *complex* system from among multiple alternatives under *uncertain* conditions.

The word "assist" emphasizes that policy analysis is used by policymakers as a decision aid, just as checklists, advisors, and horoscopes can be used as decision aids. Policy analysis is not meant to replace the judgment of the policymakers (any more than an X-ray or a blood test is meant to replace the judgment of medical doctors). Rather, the goal is to provide a better basis for the exercise of that judgment by helping to clarify the problem, presenting the alternatives, and comparing their consequences in terms of the relevant costs and benefits.

The word "complex" means that the system being studied contains so many variables, feedback loops, and interactions that it is difficult to project the consequences of a policy change. Also, the alternatives are often numerous, involving mixtures of different technologies and management policies, and producing multiple consequences that are difficult to anticipate, let alone predict.

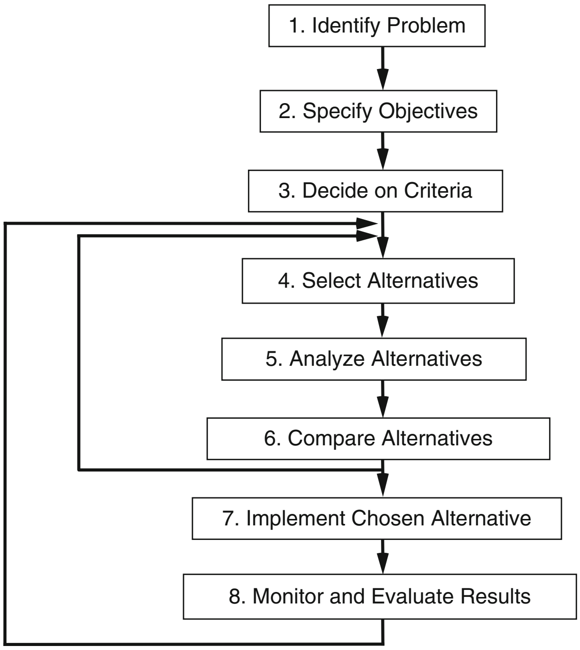
The word "uncertain" emphasizes that the choices must be made on the basis of incomplete knowledge about (a) the future world, (b) the model of the relevant system for that future world, (c) the outcomes from the system, and (d) the weights that the various stakeholders will put on the outcomes. This situation is sometimes referred to as "deep uncertainty".

Policy analysis is performed in government, at all levels; in independent policy research institutions, both for-profit and not-for-profit; and in various consulting firms. It is not a way of solving a specific problem, but is a general approach to problem solving. It is not a specific methodology, but it makes use of variety of methodologies in the context of a generic framework.

The Policy Analysis Steps

The policy analysis process generally involves performing the same set of logical steps, not always in the same order (Walker 2000; Miser and Quade 1985, p. 123). The steps, summarized in Fig. 2, are:

1. *Identify the problem.* This step sets the boundaries for what follows. It involves defining the system of interest, identifying the questions or issues involved, fixing the context within which the issues are to be analyzed and the policies will have to function (this is often done by using “scenarios”), clarifying constraints on possible courses of action, identifying the people who will be affected by the policy decision (the “stakeholders”), and discovering the major operative factors.
2. *Identify the objectives of the new policy*. Loosely speaking, a policy is a set of actions taken to solve a problem. The policymaker(s) and stakeholders have certain objectives that, if met, would “solve” the problem. In this step, the policy objectives are determined. (Most public policy problems involve multiple objectives, some of which conflict with others.)
3. *Decide on criteria (measures of performance and cost) with which to evaluate alternative policies.* Determining the degree to which a policy meets an objective involves measurement. This step involves identifying consequences of a policy that can be measured (either quantitatively or qualitatively) and that are directly related to the objectives. It also involves identifying the costs (negative benefits) that would be produced by a policy, and how they are to be measured.
4. *Select the alternative policies to be evaluated*. This step specifies the policies whose consequences are to be estimated. It is important to include as many as stand any chance of being worthwhile. If a policy is not included in this step, it will never be examined, so there is no way of knowing how good it may be. The current policy should be included as the “base case” in order to determine how much of an improvement can be expected from the other alternatives.
5. *Analyze each alternative*. This means determining the consequences that are likely to follow if the alternative is actually implemented, where the consequences are measured in terms of the criteria chosen in Step 3. This step usually involves using a model or models of the system.



Public Policy Analysis, Fig. 2 Steps in a policy analysis study (Source: Walker 2000)

6. *Compare the alternatives in terms of projected costs and benefits.* This step involves ranking the alternatives in order of desirability and choosing the one preferred. If none of the alternatives examined so far is good enough to be implemented (or if new aspects of the problem have been found, or the analysis has led to new alternatives), return to Step 4.
7. *Implement the chosen alternative*. This step involves obtaining acceptance of the new procedures (both within and outside the government), training people to use them, and performing other tasks to put the policy into effect.
8. *Monitor and evaluate the results*. This step is necessary to make sure that the policy is actually accomplishing its intended objectives. If it is not, the policy may have to be modified or a new study performed.

The individual steps in the process are described in detail by Miser and Quade (1985, Chap. 4), Quade (1989, Chap. 4), Walker (2000), and Enserink et al. (2010).

OR/MS and Public Policy Analysis

Policy analysis is closely related to operations research; in fact, in many respects it grew out of operations research as it was being applied at the RAND Corporation and other applied research organizations in the 1960s and 1970s. Miser (1980) and Majone (1985) describe this evolution. In the beginning, operations research techniques had been applied primarily to problems in which there were few parameters and a clearly defined single objective function to be optimized (e.g., aircraft design and placement of radar installations). Gradually, the problems being analyzed became broader and the contexts more complex. Health, housing, transportation, and criminal justice policies were being analyzed. Single objectives (e.g., cost minimization or single variable performance maximization) were replaced by the need to consider multiple (and conflicting) objectives (e.g., the impacts on health, the economy, and the environment and the distributional impacts on different social or economic groups). Non-quantifiable and subjective considerations had to be considered in the analysis (Schlesinger 1967, provided an early discussion of this issue). Optimization was replaced by satisficing.

Simon (1969, pp. 64–65) defined satisficing to mean finding an acceptable or satisfactory solution to a problem instead of an optimal solution. He said that satisficing was necessary because “in the real world we usually do not have a choice between satisfactory and optimal solutions, for we only rarely have a method of finding the optimum.”

Operations research techniques are among the many tools in the policy analyst’s took kit. The analyses and comparisons of alternative policies are usually carried out with the help of mathematical and statistical models. Simulation, mathematical programming, and queueing theory are among the many tools that are used in policy analysis study. But modeling is just one part of the process; all of the steps are important.

The policy analysis process has been applied to a wide variety of problems. Miser and Quade (1985, Chap. 3) provide examples of some of these, including improving blood availability and utilization, improving fire protection (for this, see also Walker et al. 1979), protecting an estuary from flooding, and providing energy for the future. More generally, the policy analysis approach has been used in the formulation of policies at the national level, including national security policies, transportation policies, and water management policies (e.g., Goeller and the PAWN Team 1985). Other examples that illustrate the approach can be found in a variety of publications, including Drake et al. (1972), House (1982), Mood (1983), Pollock et al. (1994), Miser (1995), and Walker et al. (2008).

See

- ▶ Choice Theory
- ▶ Cost Analysis
- ▶ Cost-Effectiveness Analysis
- ▶ Decision Analysis
- ▶ Decision Making and Decision Analysis
- ▶ Deep Uncertainty
- ▶ Exploratory Modeling and Analysis
- ▶ Multi-attribute Utility Theory
- ▶ Practice of Operations Research and Management Science
- ▶ RAND Corporation
- ▶ Satisficing
- ▶ Systems Analysis

References

Drake, A. W., Keeney, R. L., & Morse, P. M. (Eds.). (1972). *Analysis of public systems*. Cambridge, MA: MIT Press.

Enserink, B., Hermans, L., Kwakkel, J., Thissen, W., Koppenjan, J., & Bots, P. (2010). *Policy analysis of multi-actor systems*. Den Haag: Boom Lemma Publishers.

Findeisen, W., & Quade, E. S. (1985). The methodology of systems analysis: An introduction and overview, Chap. 4. In H. J. Miser & E. S. Quade (Eds.), *Handbook of systems analysis: Overview of uses, procedures, applications, and practice*. New York: Elsevier.

Goeller, B. F., & the PAWN Team. (1985). Planning the Netherlands’ water resources. *Interfaces, 15*(1), 3–33.

House, P. W. (1982). *The art of public policy analysis, Sage Library of Social Research* (Vol. 135). Beverly Hills, CA: Sage Publications.

Majone, G. (1985). Systems analysis: A genetic approach, Chapter 2. In H. J. Miser & E. S. Quade (Eds.), *Handbook of systems analysis: Overview of uses, procedures, applications, and practice*. New York: Elsevier.

Mayer, I. S., van Daalen, C. E., & Bots, P. W. G. (2004). Perspectives on policy analyses: A framework for understanding and design. *International Journal of Technology, Policy and Management, 4*(2), 169–190.

Miser, H. J. (1980). Operations research and systems analysis. *Science, 209*, 139–146.
Miser, H. J. (Ed.). (1995). *Handbook of systems analysis: Cases*. Chichester: Wiley.
Miser, H. J., & Quade, E. S. (Eds.). (1985). *Handbook of systems analysis: Overview of uses, procedures, applications, and practice*. Chichester: Wiley.
Mood, A. M. (1983). *Introduction to policy analysis*. New York: North-Holland.
Pollock, S. M., Rothkopf, M. H., & Barnett, A. (Eds.). (1994). *Operations research and the public sector* (Handbooks in operations research and management science, Vol. 6). New York: North-Holland.
Quade, E. S. (1989). *Analysis for public decisions*. New York: Elsevier.
Schlesinger, J. R. (1967). *On relating non-technical elements to system studies, P-3545*. Santa Monica, CA: The RAND Corporation.
Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.
Thissen, W. A. H. & Walker, W. E. (2013). Public policy analysis: new developments. New York: Springer.
Walker, W. E. (2000). Policy analysis: A systematic approach to supporting policymaking in the public sector. *Journal of Multicriteria Decision Analysis, 9*(1–3), 11–27.
Walker, W. E., Chaiken, J. M., & Ignall, E. J. (Eds.). (1979). *Fire department deployment analysis: A public policy analysis case study*. New York: Elsevier North Holland.
Walker, W. E., van Grol, R., Rahman, S. A., van de Voort, M., Röhling, W., & Burg, R. (2008). Policy analysis of sustainable transport and mobility: The SUMMA project, Chapter 13. In A. Perrels, V. Himanen, & M. Lee-Gosselin (Eds.), *Building blocks for sustainable transport: Obstacles, trends, solutions* (pp. 73–102). Bingley, UK: Emerald Group.

Pull System

Production system in which work is released into the production facility based on the current state of the facility, which includes information such as available inventory, work in process, and realized demand.

See

- ▶ CONWIP
- ▶ Kanban
- ▶ Production Management

Pure-Integer Programming Problem

A mathematical programming problem in which all variables are restricted to be integer. Usually refers to a problem in which the constraints and the objective function are linear.

See

- ▶ Mixed-Integer Programming Problem (MIP)

Push System

Production system in which work is released into the system according to forecasted demand, usually based on a schedule prepared in advance.

See

- ▶ Production Management